2004

Characterization of meiotic recombination in maize using the a1-sh2 interval as a model system

Hong Yao
Iowa State University

Characterization of meiotic recombination in maize using the *a1–sh2* interval as a model system

by

Hong Yao

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Genetics

Program of Study Committee:
Patrick S. Schnable, Major Professor
Basil J. Nikolau
Charlotte Bronson
Thomas Peterson
Volker Brendel

Iowa State University

Ames, Iowa

2004

Graduate College
Iowa State University

This is to certify that the doctoral dissertation of

Hong Yao

has met the dissertation requirements of Iowa State University

Signature was redacted for privacy.

Major Professor

Signature was redacted for privacy.

For the Major Program

TABLE OF CONTENTS

CHAPTER 1. GENERAL INTRODUCTION

Introduction

Meiotic homologous recombination refers to the DNA exchange between homologous chromosomes in the regions of high sequence similarity during meiosis. Such recombination is important to eukaryotes. It provides physical connections between pairs of homologous chromosomes to prevent non-disjunctions during meiosis. In addition, meiotic homologous recombination shuffles genetic information between the two parental haplotypes, which increases haplotype diversity and has the potential to create new alleles and thus contributes to the evolution of the organisms. Rate of meiotic recombination is used to calculate genetic distance between markers to generate genetic maps that are critical for map-guided gene cloning. Meiotic homologous recombination also shapes the genomic pattern of linkage disequilibrium (LD) (the non-random association of alleles in a population) and therefore contributes to the success of LD-based association mapping. Homologous recombination-based gene targeting is a useful tool to modify a genome via the replacement of targeted endogenous genes and other types of targeted-transgene integration. Gene targeting is very efficient for bacteria, yeast and mouse (using embryonic stem cells) but not for higher plants. The poor efficiency of gene targeting in higher plants is due to the very low rate of homologous recombination in plant somatic cells. Hence, strategies to enhance homologous recombination have the potential to improve gene-targeting techniques for higher plants. Since homologous recombination is very efficient during meiosis in plants, study of meiotic homologous recombination can reveal factors that affect the efficiency of homologous recombination and contribute to the improvement of gene-targeting techniques. This study also contributes to the development of other tools and strategies for crop improvement such as the introgression of useful genes from one species into another. Thus, the study of the mechanisms of meiotic homologous recombination will not only enable us to

better understand this fundamental cellular process itself, but will also help us to develop and use crop improvement techniques that are based on this process.

Detailed studies of meiotic homologous recombination in *S. cerevisiae* have revealed many aspects of the mechanisms of this process. Nonetheless, in higher eukaryotes such as plants, the mechanisms of meiotic homologous recombination have not been well characterized due to a shortage of techniques to study meiotic homologous recombination in organisms with complex genomes. In most eukaryotes, meiotic recombination does not occur randomly along chromosomes (reviewed by LICHTEN and GOLDMAN 1995), which reflects that genetic distances are poorly correlated with physical distances. Identifying the reasons for the non-uniform distribution of meiotic homologous recombination events can be the first step in unveiling the mechanisms of meiotic homologous recombination in higher eukaryotes. In this study, the *a1-sh2* interval was used as a model for studying meiotic homologous recombination in maize. The size of this interval is small enough for relatively easy cloning (CIVARDI *et al.* 1994) and further physical characterization. The *a1* and *sh2* loci that flank this interval confer visible kernel phenotypes. This makes it easy to identify recombinants by scoring large populations. Rates and distribution of meiotic homologous recombination events were determined across the *a1-sh2* interval. This could reflect features of meiotic homologous recombination along the chromosomes and across the genome. The long-term goal of this research is to answer the question, "Why does meiotic recombination occur non-randomly in the maize genome?" The specific objectives of this research were to characterize how meiotic homologous recombination events are distributed relative to genes and non-genic regions across the *a1-sh2* interval and to characterize the influence of genetic *cis*-modifiers on the rate of recombination and the distribution of recombination breakpoints across this interval.

Dissertation Organization

This dissertation consists of three journal papers (chapters 2 to 4) and a chapter of general conclusions (chapter 5). The paper in chapter 2 has been published in PNAS. My contributions to this paper included partial sequencing the *A1 Sh2* haplotype from the LH82, *a1::rdt sh2*, and Line C stocks, identifying the *yz1* gene via bioinformatics and molecular analyses, contributing to the identification and characterization of the *x1* gene, developing the RFLP markers at the *x1* locus and mapping recombination breakpoints using the *x1*-probe, characterization of the Interloop Region, mapping the recombination breakpoints across the Interloop Region, *yz1* and the *x1* loci to high resolution, double-checking and, in some instances, experimentally validating mapping results obtained by previous graduate students, remapping eight recombination breakpoints that had been inadvertently placed in the wrong location thereby establishing that the region between *a2* and *sh2* loci is a recombination cold spot. Moreover, I made major contribution to the writing of the paper under the guidance of Dr. Patrick S. Schnable, my major professor. The paper in chapter 3 is to be submitted to Genetics. I did all the research reported in this manuscript and wrote most of the paper under the guidance of Dr. Schnable. The paper in chapter 4 is to be submitted to Plant Molecular Biology. My contribution to this work included creating and characterizing data set 1 derived from sequences of eight maize genes cloned in Dr. Schnable's lab, evaluating the seven gene prediction programs using data set 1, identifying and analyzing gene structures of the *yz1* and *x1* genes, developing strategies to create data set 2 and to evaluate FGENESH, GENSCAN and GeneMark.hmm using data set 2 with Ling Guo, Yan Fu, Dr. Ashlock and Dr. Schnable. I also wrote the majority of the paper under Dr. Schnable's guidance. The General Conclusions section of this dissertation summarize and discuss the overall results relative to the specific questions addressed in the General Introduction.

Abbreviations

DSB, double strand breaks; HR, homologous recombination; NHEJ, non-homology end-joining; DSBR, double-strand break repair, SDSA, synthesis-dependent strand annealing; SSA, single-strand annealing; BIR, break-induced replication; CO, crossover; NCO, non-crossover; DHJ, double Holliday junctions; hDNA, heteroduplex DNA; InDels, insertion/deletion polymorphisms; SNPs, single nucleotide polymorphisms; LD, linkage disequilibrium; MMS, methyl methanesulfonate.

Literature Review

Pathways to repair double strand breaks (DSB)

DNA double strand breaks (DSB) are deleterious to a cell. As such, they need repairing to prevent possible lethal effects on the cell. DSB can be repaired via several mechanisms that can be classified as homologous recombination (HR) pathways and non-homology end-joining (NHEJ) pathways. HR occurs between DNA fragments that are identical or of high sequence similarity over hundreds of base pairs. Templates used to repair a DSB in a DNA molecule may be its allelic sequence from homologous chromosome or sister chromatid; such a pathway is usually conservative. Yet, DSB repair may also use templates from ectopic homologous sequences that reside on the same or other chromosomes. As such mechanisms of DSB repair could result in chromosome rearrangements such as deletions, insertions, inversions and translocations, they are more likely to be mutagenic. The major mechanisms for HR include those that result in crossover (CO) or non-crossover (NCO) events via pathways proposed in the double-strand break repair model (DSBR) and the synthesis-dependent strand annealing (SDSA) model, single-strand annealing (SSA) and break-induced replication (BIR) (Figure 1) (reviewed by PAQUES and HABER 1999). In contrast to HR, NHEJ does not require sequences to share long stretches of homology. The two ends of a DSB can be ligated together without losing any genetic information (Figure 1b) but such events are rare. More frequently, during NHEJ the

ends of a DSB are processed. Microhomology (usually less than 5 bp) between sequences flanking the processed ends may help stabilize the intermediate product. The DNA may thus be repaired via a SSA-like process (Figure 1d).

DSBs can happen anytime during the lifetime of a cell (reviewed by KUPIEC 2000) and can be induced directly or indirectly by exogenous DNA damaging agents (reviewed by PUCHTA and HOHN 1996; BRITT 1999; GORBUNOVA and LEVY 1999; KUPIEC 2000; VAN GENT *et al.* 2001; VAN DEN BOSCH *et al.* 2002). These agents include irradiation such as X-rays or ionizing radiation and chemicals such as bleomycine or methyl methanesulfonate (MMS). DSBs can also arise "spontaneously" via the action of endogenous factors. These endogenous factors include free radicals generated from normal metabolic pathways, nicks on a single-stranded DNA during DNA replication, collapsed replication forks, mechanical stress on chromosomes, and transpositions of transposons. DSBs also occur during programmed events in a cell such as meiotic recombination, the switching of mating-type genes in yeast, and the V(D)J recombination in mammalian cells.

Mechanisms that are chosen by a cell to repair DSBs are organism-dependent and cell-cycle-dependent. Both HR and NHEJ pathways are important to a cell. In yeast, however, HR is preferred over NHEJ in a mitotic cell, whereas in higher eukaryotes such as plants and mammals NHEJ is preferred. During the S and G2 phases of the cell cycle, the efficiency of HR is increased since DSB repair can use sister chromatids as templates. In the G1 phase, NHEJ may be more efficient since sister chromatids are not available (reviewed by VAN GENT *et al.* 2001). During meiosis, HR is dominant.

Models of meiotic homologous recombination and supporting evidence

In addition to the repair of DSBs to maintain the integrity of a genome, meiotic HR has unique importance to a cell. It provides physical connections between homologous chromosomes to prevent chromosome disjunction during meiosis. It also generates allelic diversity upon which selection can act. The two products from meiotic recombination (CO

and NCO) are usually associated with each other, which can be explained by the canonical DSBR model for recombination (SZOSTAK *et al.* 1983; CAO *et al.* 1990; SUN *et al.* 1991). Under this model (Figure 1), meiotic recombination is initiated by a DSB. Following the 5' to 3' resection at both ends of the DSB, one of the resulting 3' single-stranded overhangs can invade the duplex DNA of its homologue to form a D-loop. DNA synthesis primed from the invading 3' single-stranded overhang enlarges the D-loop, which causes the capture of the other 3' overhang. Ligation of the newly synthesized DNA primed from the invading 3' overhang and the resected 5' end form double Holliday junctions (DHJ). Resolution of the DHJ by cutting the DNA strands at each HJ in the same or different manners can result in either NCO (e.g., gene conversion without CO) or CO (e.g., gene conversion with CO of the flanking markers). Several lines of evidence strongly support the DSBR model. DSBs have been observed to be associated with several hot spots for meiotic recombination during meiosis (SUN *et al.* 1989; CAO *et al.* 1990; BULLARD *et al.* 1996; reviewed by LICHTEN and GOLDMAN 1995; PETES 2001). The time courses of these DSBs are consistent with the expected kinetics under the assumption that these DSBs initiate recombination events at these hot spots (PADMORE *et al.* 1991; GOYON and LICHTEN 1993). In addition, the frequency and distribution of DSBs are consistent with the frequency and distribution of meiotic recombination events in the yeast genome (BAUDAT and NICOLAS 1997; WU and LICHTEN 1994). Resection of the DSB ends to expose 3' single-stranded overhangs has been demonstrated in several physical studies (CAO *et al.* 1990; SUN *et al.* 1991; BISHOP *et al.* 1992). The recombination intermediate, the DHJ, has also been observed (COLLINS and NEWLON 1994; SCHWACHA and KLECKNER 1994, 1995).

Some aspects of the DSBR model, however, are not consistent with experimental observations. First, the DSBR model predicts that the ratio of gene conversion with CO of flanking markers to gene conversion without CO will be 1:1. This is based on the model's prediction that CO and NCO should result from resolution of the DHJ with equal chance. In

contrast, in yeast, only 35% of meiotic gene conversion events are associated with CO of the flanking markers and the frequency of such events in mitosis is even lower (10 to 20%) (reviewed by PRADO and AGUILERA 2003; VAN DEN BOSCH *et al.* 2002). Second, the DSBR model predicts that the formation of heteroduplex DNA (hDNA) in both the donor (containing the intact DNA duplex) and the recipient (containing the DSB) loci, which will end up in different chromatids, and that gene conversion events that result from the repair of mismatches in hDNA should be located at both sides of the DSB. Yet, genetic studies of meiotic recombination at the *HIS4* and *ARG4* loci showed that hDNAs and conversions are most frequently located at only one side of the DSB and that when hDNAs can be detected on both sides of the DSB, they are located on the same chromatid (PORTER *et al.* 1993; GILBERTSON and STAHL 1996). Third, under the DSBR model, the hDNA is formed before the DHJ. Experimental hDNA, however, can not be detected until around the time of DHJ resolution (NAG and PETES, 1993; SCHWACHA and KLECKNER 1995). To explain these results, an alternative model, SDSA model was proposed (reviewed by PAQUES and HABER 1999). In one of the widely accepted versions of the SDSA model, recombination initiation and the resection of the DSB ends are the same as those proposed in the DSBR model (Figure 1). One of the 3' overhangs invades its homologous duplex and primes DNA synthesis. Although the D-loop is formed, it is not enlarged. Hence, it cannot be captured by the other 3' overhang. DNA synthesis drives the migration of the D-loop. The newly synthesized DNA is displaced and is eventually captured by the other DSB end at the recipient locus. The single-stranded gap is filled either via DNA synthesis primed from the non-invading 3' overhang or via the lagging-strand DNA synthesis coupled to the leading-strand DNA synthesis primed from the invading 3' overhang. Only NCO events result and hDNA is present only at the recipient locus. Recently, ALLERS and LICHTEN (2001) found that hDNAs in NCO events and DHJs are formed at the same time, whereas CO events appear at the time when DHJs are resolved. In addition, the *ndt80* mutant (Ndt80p is a

meiosis-specific transcription factor) causes the failure of DHJ resolution and reduces the frequency of CO. NCO events, however, are not affected by the mutation (ALLERS and Lichten 2001). These results suggest that meiotic CO and NCO events are produced from different pathways. As proposed by ALLERS and LICHTEN (2001), CO events result from the DSBR pathway and NCO events are from the SDSA pathway. Identification of another branched recombination intermediate, the single-end invasion, and the kinetic study of this intermediate relative to other events during meiotic recombination are also consistent with this view (Hunter and KLECKNER, 2001).

Proteins involved in the meiotic recombination pathways and their plant homologues

Models for meiotic recombination are proposed based on results from extensive genetic and physical studies in yeast. Many genes responsible for different steps in meiotic recombination have been identified and characterized genetically and molecularly in yeast. Many of these genes are conserved in higher eukaryotes such as mammals and plants.

Initiation of meiotic recombination. Meiotic recombination is initiated by DSBs of which the induction needs at least 11 different genes in yeast (*RAD50*, *SPO11*, *MRE11*, *XRS2*, *MEI4*, *MER1*, *MER2*, *MRE2*, *REC102*, *REC104* and *REC114*, Figure 1) (reviewed by PAQUES and HABER 1999). *RAD50*, *MRE11* and *XRS2* also play roles in mitotic recombination whereas the other eight genes are specific for meiotic recombination (reviewed by SMITH and NICOLAS 1998). Null mutants of any of these genes almost abolish meiotic recombination because the induction of DSBs is defective in these mutants. Products of the *SPO11* gene are covalently bound to the 5' ends of meiotic DSBs in the *rad50s* strain that accumulates unresected DSBs (KEENEY *et al.* 1997). Spo11p has homology to Type II topoisomerase from archaebacteria (BERGERAT *et al.* 1997). These findings suggest that Spo11p is the endonuclease that generates DSBs via catalyzing a transesterification reaction. Homologues of the yeast SPO11 gene have been found in *S. pombe*, *C. elegans*, *D. melanogaster*, mouse, human and Arabidopsis (reviewed by PAQUES and HABER 1999;

COHEN and POLLARD 2001; SCHWARZACHER 2003). In contrast to other organisms in which there is only one homologue of the yeast *SPO11* gene, Arabidopsis contains three homologues of *SPO11* (reviewed by SCHWARZACHER 2003). The *AtSPO11-1* gene functions during meiosis. Homozygous mutants of the *AtSPO11-1* exhibit aberrant meiosis, disrupted synapsis and reduced meiotic recombination (GRELON *et al.* 2001). These findings suggest that the function of the Spo11 protein to initiate meiotic recombination may be conserved in plants. The biological functions of the *AtSPO11-2* and *AtSPO11-3* are not clear. A recent study demonstrated that the *AtSPO11-3* gene is important to plant growth and somatic development (YIN *et al.* 2002). Specific functions of most the other ten yeast genes in inducing meiotic DSBs are still not clear except for the functions of *MER1* and *MRE2* which are needed for the meiotic specific splicing of the transcript from the *MER2* gene (reviewed by PAQUES and HABER 1999). Three additional genes, *RED1*, *HOP1* and *MEK1/MRE4*, also play roles in the regulation of the level of meiotic DSBs. Although mutations of these three genes do not abolish the meiotic induction of DSBs, they do reduce the level of meiotic DSBs (reviewed by PAQUES and HABER 1999). There are two homologues of the *HOP1* gene in Arabidopsis, *ASY1* and *ASY2* (reviewed by SCHWARZACHER 2003). The *ASY1* gene functions in meiosis and is essential for normal synapsis (CARYL *et al.* 2000; ARMSTRONG *et al.* 2002).

Resection of DSB ends. DSB resection requires the removal of the Spo11p from the 5' ends of DSB and 5' to 3' degradation of the 5' ends to expose the 3' single-strand tails. Evidence from yeast suggests that *RAD50*, *MRE11* and *COM1/SAE2* are involved in the DSB end resection during meiotic recombination (reviewed by PAQUES and HABER 1999). Null mutants of *RAD50* and *MRE11* abolish the induction of meiotic DSBs whereas some separation-of-function (hypermorphic) mutations of *RAD50* and *MRE11* allow the induction of meiotic DSBs but prevent the processing of DSB ends (reviewed by PAQUES and HABER 1999). The yeast Mre11p has nuclease activities, which are stimulated by the Rad50p

(reviewed by SYMINGTON 2002). These two proteins as well as the *XRS2* protein form a complex that is involved in many other cellular processes (e.g., mitotic recombination, telomere maintenance etc.) in addition to meiotic recombination (reviewed by SYMINGTON 2002). Meiotic DSB induction also fails in the null mutant of *XRS2* but no separation-of-function mutations of *XRS2* have been identified that accumulate unprocessed meiotic DSBs. Since, however, the Xrs2p works with Rad50p and Mre11p in the 5' to 3' resection of the DSB ends in mitosis, the Xrs2 protein may play roles in the resection of meiotic DSB ends too. Null mutation of *COM1/SAE2* exhibits the same phenotype as the separation-of-function mutations of *RAD50* and *MRE11*, in which Spo11p cannot be removed from the 5' ends of DSB (reviewed by PAQUES and HABER 1999). Homologues of the *RAD50* and *MRE11* genes have been identified in plants; both are single-copy in the Arabidopsis genome (reviewed by SCHWARZACHER 2003). The RAD50 and MRE11 proteins have been found to form a complex in Arabidopsis cells (DAOUDAL-COTTERELL *et al.* 2002). The Arabidopsis *rad50* mutant is sterile and exhibits somatic hyper-homologous recombination and MMS sensitivity (GHERBI *et al.* 2001; GALLEGO *et al.* 2001). Similarly, the Arabidopsis *mre11* mutants show sensitivity to MMS and X-ray; mutants at the 5' conserved region of the MRE11 protein are sterile (BUNDOCK and HOOYKAAS 2002). These results suggest that, like their yeast homologues, the Arabidopsis RAD50 and MRE11 proteins may play important roles in meiotic and mitotic recombination. So far no homologues of the yeast *XRS2* and *COM1/SAE2* have been found in plants, which suggests that plants may have some unique features in the resection of the DSB ends.

Strand invasion. In *S. cerevisiae* after resection of the meiotic DSBs the invasion of the exposed 3' tail to its homologue duplex requires the functions of eleven genes: *RAD51*, *RAD52*, *RAD54*, *RAD55*, *RAD57*, *RDH54/TID1*, *DMC1*, *SAE3*, *RFA1* (reviewed by SMITH and NICOLAS 1998), *MER3* (NAKAGAWA and OGAWA 1999) and *MND1* (GERTON and DERISI 2002). The *RAD51*, *RAD55*, *RAD57* and *DMC1* genes are homologues of the *recA* gene

from E. *coli*. The RECA protein stimulates strand exchange during homologous recombination (reviewed by SYMINGTON 2002). The *RAD51*, *RAD55*, *RAD57* belong to the *RAD52* epistasis group, which also includes *RAD52*, *RAD54*, *RDH54/TID1*, *RAD50*, *MRE11* and *XRS2* (reviewed by SYMINGTON 2002). Genes in the *RAD52* epistasis group also play important roles in mitotic homologous recombination and mutants of this group of genes are sensitive to the DNA damage caused by ionizing radiation (reviewed by SYMINGTON 2002). Although a homologue of *recA*, *DMC1* does not belong to the *RAD52* epistasis group because its mutant is not sensitive to ionizing radiation (reviewed by SYMINGTON 2002). The *RFA1* gene encodes a subunit of the RPA protein complex that binds ssDNA and has been suggested to play a role in removal of secondary structures to assist the formation of uninterrupted Rad51 nucleoprotein filament (reviewed by SYMINGTON 2002). Mutations of the *RAD51*, *RAD55*, *RAD57*, *DMC1*, *RAD52* and *RFA1* exhibit defective meiotic recombination phenotypes such as the accumulation of hyperresected 5' ends of DSBs, reduced levels of DHJ formation and crossover products (reviewed by SYMINGTON 2002). Biochemical studies suggest Rad51p plays a major role in strand invasion with the assistance of several other proteins. The proposed mechanism is that Rad52p and a heterodimer of Rad55p and Rad57p work together to displace the RPA complex from the 3' single-strand tail so that the Rad51p can bind the 3' single-strand tail to form nucleoprotein filament and stimulate strand invasion (reviewed by SMITH and NICOLAS 1998). The specific role of Dmc1p in the strand invasion is not clear. Double mutants of *DMC1* and *RAD51* exhibit a more reduced level of CO events than either single mutant (reviewed by SYMINGTON 2002). Cytogenetic analyses have revealed that during early meiotic prophase Dmc1 and Rad51 foci co-localize to a significant degree and both may be components of the early recombination nodules, structures that can be detected by electron microscopy during leptotene and zygotene (reviewed by PAQUES and HABER 1999). The formation of Dmc1 foci requires Rad51p, in the absence of Dmc1p the dissociation of Rad51 foci fails (reviewed by SMITH

and NICOLAS 1998; PAQUES and HABER 1999). No direct interactions, however, were detected between the Dmc1p and Rad51p, Rad52p or Rad54p (DRESSER *et al.* 1997). These results suggest Dmc1p and Rad51p have overlapping but different roles in meiotic recombination. It has been suggested that Dmc1p may function specifically in recombination between homologues during meiosis (SCHWACHA and KLECKNER 1997). The Rdh54/Tid1 protein may also be involved in the same pathway as Dmc1p to promote meiotic interhomologue recombination (reviewed by PAQUES and HABER 1999). It has been hypothesized that Rdh54p and its paralogue Rad54p function to promote D-loop formation (reviewed by VAN DEN BOSCH *et al.* 2002). The *rdh54/tid1 rad54* double mutant exhibits more severe defects in meiotic recombination than either single mutant that affects meiotic recombination only moderately (reviewed by SYMINGTON 2002). *SAE3* may also function in the same pathway as *DMC1* because the expression of both genes is induced at same time and the mutant phenotypes of *sae3*, *dmc1* and *sae3 dmc1* double mutants are the same (reviewed by SMITH and NICOLAS 1998). The *RED1* and *HOP1* gene products may also help to channel meiotic DSBs to be repaired by the interhomologue recombination pathway in which Dmc1p is involved (reviewed by SMITH and NICOLAS 1998; PAQUES and HABER 1999). The *MER3* gene that encodes a meiosis-specific DNA helicase may also be involved in strand invasion. The *mer3* mutant accumulates a fraction of DSBs that are hyperresected late in meiosis (NAKAGAWA and OGAWA 1999). The Mnd1 protein is also meiosis-specific; mutation of the *MND1* gene abolishes the formation of Holliday junctions (joint molecules) and COs and causes the DSB ends to be slightly hyperresected but the initiation of DSBs is not affected (GERTON and DERISI, 2002).

Homologues of several key proteins involved in strand invasion have been identified in plants (reviewed by BHATT *et al.* 2001; SCHWARZACHER 2003). Characterization of the *rad51* and *dmc1/lim15* genes suggests that their functions in meiotic recombination are similar to their yeast homologues (reviewed by BHATT *et al.* 2001; SCHWARZACHER 2003;

VERGUNST and HOOYKAAS 1999). Mutation of the single-copy *dmc1* gene of Arabidopsis dramatically reduces the plant fertility, which is correlated with abnormal female and male meiosis (COUTEAU *et al.* 1999). Two homologues of *RAD51* exist in maize, *rad51A* and *rad51B*. Homozygous plants with double mutant *rad51a* and *rad51b* alleles are male sterile and show reduced female fertility but are viable (J. LI and P. S. SCHNABLE, manuscript in preparation). Cytogenetic study revealed that these mutant phenotypes are correlated with abnormal chromosome behavior during meiosis (J. LI and P. S. SCHNABLE, manuscript in preparation). Homologues of *RAD54*, *RDH54/TID1*, *RAD57*, *RFA1* and *MND1* are also present in plants (reviewed by SCHWARZACHER 2003; http://weilthing.agry.purdue.edu/recgenesXL.html). Although *RAD52* is critical to homologous recombination in yeast, no plant homologues of *RAD52* have been identified so far.

Correction of mismatches in hDNA. Current models for meiotic recombination agree that gene conversion events result from the correction of mismatches in the hDNA. In yeast, five genes are involved in mismatch repair in meiotic recombination: *MSH2*, *MSH3*, *MSH6*, *PMS1*, and *MLH1* (reviewed by PAQUES and HABER 1999). Msh2p, Msh3p and Msh6p are homologues of the MutS protein from *E. coli* (reviewed by BORTS *et al.* 2000). Msh2p and Msh3p form a heterodimer and bind mispaired nucleotides; another heterodimer consisting of Msh2p and Msh6p binds small insertion/deletion polymorphisms (InDels) (reviewed by PAQUES and HABER 1999). Repair of these mismatches also requires the function of the heterodimer consisting of Pms1p and Mlh1p, two homologues of the MutL protein from *E. coli* (reviewed by PAQUES and HABER 1999). Plant homologues of these mismatch repair proteins have been identified (reviewed by BHATT *et al.* 2001). Unlike in yeast there are two MSH6-like proteins in Arabidopsis, AtMSH6 and AtMSH7 (CULLIGAN and HAYS 2000). In vitro studies showed that the interactions between these MSH proteins and the substrate specificities of the AtMSH2/AtMSH3 complex and the AtMSH2/AtMSH6

complex are similar to those in yeast. The substrate specificity of the AtMSH2/AtMSH7 seems different from others (CULLIGAN and HAYS 2000), however. Inactivation of AtMSH2 in Arabidopsis does not cause sterility but reduces the stability of repeat sequences in the genome (LEONARD *et al.* 2003).

Regulation of CO. Meiotic recombination results in two products, CO and NCO, which are regulated differently. The distribution of COs is controlled uniquely in that it exhibits CO interference (a CO at one site on a chromosome reduces the chances of another nearby CO occurring), obligate chiasma (each chromosome has at least one CO), and regulation by size (CO frequency is negatively associated with chromosome size) (reviewed by PAQUES and HABER 1999). In yeast, several genes are involved in the control of CO interference: *ZIP1*, *MER3*, *MSH4*, *NDJ1/TAM1*, *RDH54/TID1*, and *DMC1* (reviewed by PAQUES and HABER 1999; NAKAGAWA and OGAWA 1999; SHINOHARA *et al.* 2003). Zip1p is a component of the central region of the synaptonemal complex (SC), a protein structure formed between homologous chromosomes when they synapse (reviewed by PAQUES and HABER 1999). *MER3* encodes a DNA helicase and the splicing of its transcript is controlled by *MER1* and *MRE2* (reviewed by PAQUES and HABER 1999; NAKAGAWA and OGAWA 1999). Msh4p is a homologue of MutS, a mismatch repair protein in *E. coli*; yet, Msh4p is not involved in mismatch repair in yeast (reviewed by BORTS *et al.* 2000). *NDJ1/TAM1* encodes a telomere-associated protein (CHUA and ROEDER 1997). The *dmc1* mutant abolishes meiotic recombination. Its role in regulating CO interference is suggested by a study using a *dmc1* mutant strain in which Rad54p was over expressed (SHINOHARA *et al.* 2003). Mutations of *NDJ1/TAM1*, *RDH54/TID1*, and *DMC1* (with over expressed Rad54p) abolish CO interference but the frequency of COs is normal. In contrast, mutations of *ZIP1*, *MER3* and *MSH4* reduce both CO interference and CO frequency (reviewed by PAQUES and HABER 1999; NAKAGAWA and OGAWA 1999; SHINOHARA *et al.* 2003). Additional genes that affect CO frequency include *MSH5*, *MLH1*, *ZIP2* and *SGS1* (reviewed by PAQUES and

HABER 1999; LOUIS and BORTS 2003). Like Msh4p, Msh5p is a Muts homologue but is not involved in mismatch repair (reviewed by BORTS *et al.* 2000). Zip2p is also a component of the SC, which suggests the SC not only affects CO interference but also CO frequency (reviewed by PAQUES and HABER 1999). Unlike other genes, a non-null mutation of the *SGS1* that encodes the RecQ helicase increases the frequency of CO (ROCKMILL *et al.* 2003). Hence, it has been suggested that *SGS1* limits COs by helping to determine whether an initiated recombination event resolves as a CO or NCO (reviewed by LOUIS and BORTS 2003). With the exception of the *DMC1*, *MLH1*, *SGS1* and *RDH54/TID1* (HARTUNG *et al.* 2000; BAGHERIEH-NAJJAR *et al.* 2003; http://weilthing.agry.purdue.edu/recgenesXL.html) genes, plant homologues of the remaining yeast genes involved in the regulation of CO have not been reported.

Resolution of DHJs. The eukaryotic DHJ resolvase has puzzled researchers for a long time. The RuvABC complex, a Holliday junction resolvase, has been identified in bacteria; it can resolve DHJs in ways predicted by the DSBR model (WEST 1996, 1997). Although current models for meiotic recombination predict that COs arise from the resolution of DHJs, in *S. pombe*, CO may be generated solely by a Mus81 endonuclease catalyzed pathway, in which no DHJs are formed and no CO interference is involved (reviewed by HOLLINGSWORTH and BRILL 2004). In *S. cerevisiae*, genetic and physical data showed that the pathway involving DHJ and CO interference is the major pathway to generate COs and that although Mus81 has HJ resolvase activity, it is not the major resolvase (reviewed by HEYER *et al.* 2003). Recently, it was demonstrated that in mammalian cells the HJ resolvase with associated branch migration activity involves Rad51C and Xrcc3 proteins (LIU *et al.* 2004). This resolvase functions in the same way as the bacteria RuvABC resolvase (LIU *et al.* 2004). This finding opens a new door for studies to identify the DHJ resolvase in *S. cerevisiae* and higher eukaryotes. In Arabidopsis, both the *ruvB*-like gene (*AtRUVB*) and the *Rad51*-like genes (*AtRAD51C* and *AtXRCC3*) have been identified

(reviewed by BHATT *et al.* 2001; OSAKABE *et al.* 2002), which suggests the mechanism to generate COs may be similar to that in yeast and mammals. In addition, the yeast *NDT80* and *CDC5* genes are also required for resolution of DHJs to form COs (ALLERS and LICHTEN 2001; CLYNE *et al.* 2003). *NDT80* encodes a transcription factor that activates expression of many genes during meiosis (CHU *et al.* 1998; HEPWORTH *et al.* 1998) and *CDC5* encodes a kinase, the expression of which is activated by the Ndt80p. No plant homologues of the *NDT80* and *CDC5* have been reported so far.

Meiotic recombination occurs non-randomly in a genome.

Meiotic recombination does not occur randomly across eukaryotic genomes (reviewed by LICHTEN and GOLDMAN 1995). There are hot and cold spots for meiotic recombination that are regions exhibiting rates of recombination much higher and lower, respectively than the genome's average. The non-randomness of the distribution of meiotic recombination events in a genome is reflected at different levels. At the chromosomal level, genetic distances between markers across a chromosome are not correlated with the corresponding physical distances (reviewed by PUCHTA and HOHN 1996; SCHNABLE *et al.* 1998; DE MASSY 2003; SCHWARZACHER 2003). Molecular and cytological studies revealed that meiotic recombination is suppressed in regions surrounding the centromeres in many organisms such as *S. cerevisiae* (LAMBIE and ROEDER 1988), Drosophila (reviewed by KOROL *et al.* 1994), human (reviewed by LICHTEN and GOLDMAN 1995; PETES 2001; INTERNATIONAL HUMAN GENOME SEQUENCING CONSORTIUM 2001), Arabidopsis (SCHMIDT 1995; COPENHAVER *et al.* 1998; HAUPT *et al.* 2001), tomato (TANKSLEY *et al.* 1992; FRARY *et al.* 1996), maize (ANDERSON *et al.* 2003) and wheat (WERNER *et al.* 1992; GILL *et al.* 1996a, b). On the other hand, rates of recombination are elevated in regions near the telomeres in human, tomato, maize and wheat. Based on the fact that physical sizes of eukaryotic genomes are very different whereas the lengths of total genetic maps are fairly constant and on the assumption that the numbers of genes are also fairly constant among

different eukaryotes, a prediction that is now supported by the results from genome sequencing projects, THURIAUX (1977) proposed that meiotic recombination may be confined to genes. Consistent with this hypothesis, gene-rich regions are more recombinationally active than gene-poor regions in wheat (GILL *et al.* 1996a, b; FARIS *et al.* 2000) and barley (KUNZEL *et al.* 2000). In addition, maize genes tend to be recombination hot spots (reviewed by LICHTEN and GOLDMAN 1995; SCHNABLE *et al.* 1998). Recombination activities can differ even within genic hot spots. In *S. cerevisiae*, gene conversions at several loci exhibit a polarity gradient. That is, the chance that a marker will be converted decreases with its distance from the initiation site and markers that are converted at low frequencies are often co-converted with markers that are converted at high frequencies (reviewed by PAQUES and HABER 1999). At some maize loci, meiotic intragenic recombination events are clustered to either 5' or 3' ends of the genes (reviewed by SCHNABLE *et al.* 1998).

In yeast *S. cerevisiae*, the non-random distribution of meiotic recombination is correlated with the non-random distribution of DSB sites (reviewed by LICHTEN and GOLDMAN 1995; DE MASSY 2003). This suggests that the molecular basis for the non–randomness of meiotic recombination is determined via the regulation of its initiation. The level and distribution of DSBs are usually characterized in strains that carry the non-null mutant *rad50S*, which results in accumulation of DSBs and interrupts later steps of meiotic recombination (ALANI *et al.* 1990; CAO *et al.* 1990). Consistent with the nonrandom distribution of meiotic recombination at different levels, distribution of DSBs is regulated at different levels too. At the chromosomal level, large "hot" domains with high levels of DSBs are interspersed with large "cold" domains with low levels of DSBs; the differences in the rates of DSBs between the hot and cold domains can be as large as 50 fold (BAUDAT and NICOLAS 1997; GERTON *et al.* 2000). Some of the cold domains are located in telomeres and near centromeres (BAUDAT and NICOLAS 1997; GERTON *et al.* 2000). These observations suggest that locations of DSBs can be regulated via gross chromosome structures. The

distribution of DSBs can also be regulated at the local level via region-specific chromatin structures and/or DNA sequences. Most DSB sites are located in regions that are promoters or intergenic regions near the binding sites for transcription factors (BAUDAT and NICOLAS 1997; GERTON *et al.* 2000; reviewed by PETES 2001). Even so, transcription activity is not needed for DSB formation (reviewed by PETES 2001). DSB sites are hypersensitive to DnaseI and micrococcal nuclease and some exhibit such sensitivities that are induced in meiosis and that are correlated well with changes in DSB hot-spot activities (reviewed by LICHTEN and GOLDMAN 1995; PETES 2001). Some DSB hot spots have been created via the integration of foreign DNA fragments into the yeast genome. This may result in nucleosome-excluding regions. These DSB sites also show hypersensitivities to nuclease (reviewed by LICHTEN and GOLDMAN 1995; PETES 2001). Based on these results, it has been hypothesized that DSB hot sites have open chromatin structures that are accessible to the recombination machinery. Open chromatin structure, however, is not sufficient to induce high levels of meiotic DSBs since not all regions that are hypersensitive to nuclease are hot spots for DSB formation during meiosis (reviewed by LICHTEN and GOLDMAN 1995). DSB sites are not sequence specific. No consensus sequence motif that is conserved among yeast hot spots has been identified (reviewed by PETES 2001). Rather, within a DSB hot spot, DSBs can occur anywhere in a region of 70-250 bp (reviewed by DE MASSY 2003). Yet, mutations of sequences at some DSB sites can change the DSB levels and recombination activities (reviewed by LICHTEN and GOLDMAN 1995; HARING *et al.* 2003). Distribution of DSB hot spots is also positively correlated with the GC content in the yeast genome (GERTON *et al.* 2000). In addition, premeiotic DNA replication seems to have local effects on DSB formation (reviewed by PETES 2001; MURAKAMI *et al.* 2003), which suggests a correlation among replication, meiosis-specific chromatin transition and DSB formation.

In higher eukaryotes, the molecular basis for the non-random distribution of meiotic recombination is not clear. Given the conservation of many key proteins such as Spo11p

between lower and higher eukaryotes (reviewed by BHATT *et al.* 2001; COHEN and POLLARD 2001; VAN DEN BOSCH *et al.* 2002; SCHWARZACHER 2003), many aspects of the mechanisms for meiotic recombination such as initiation by DSBs may be conserved among them. It is likely that the distribution of meiotic recombination events is also realized via regulation of the initiation sites, the DSBs. Results from high-resolution mapping of the CO breakpoints and gene conversion tracts across large chromosome intervals as well as small recombination hot spots in mouse and human support this view (reviewed by DE MASSY 2003; KAUPPI *et al.* 2004). Recently, using a PCR-based method to detect meiotic DSBs the correlation between frequencies and locations of meiotic DSBs and COs has been demonstrated at the mouse *H2-Ea* recombination hot spot (QIN *et al.* 2004). In plants, regions with hyper-recombination activity across a chromosome are associated with high gene density (reviewed by SCHNABLE *et al.* 1998). Are genes per se recombinationally hyperactive and are other regions recombinationally inert simply because they lack genic recombination hot spots? Alternatively, perhaps what matters are structural features associated with high gene density, i.e., high gene density can create a recombination hyperactive domain that is larger than a gene. In the region surrounding the maize *bz1* loci, the 108-kb gene-poor but retrotransposon-rich proximal portion exhibits a recombination rate that is lower by two orders of magnitude than the 12-kb gene-rich but retrotransposon-absent distal portion (FU *et al.* 2002). Because the retrotransposon-rich portion is hypermethylated and the gene-rich portion is hypomethylated, it has been suggested that the corresponding chromosome structures may differ resulting in these dramatic different recombinational activities (FU *et al.* 2002). These results suggest that meiotic recombination in plants could be regulated via gross chromosome structures associated with high gene density. This study, however, could not map recombination breakpoints to high resolution due to the lack of polymorphic markers. Hence, the questions of whether all genes contribute significantly to high recombination activity in gene-rich regions and whether all non-genic sequences especially

the retrotransposons are recombination inert remain open. Characterization of intragenic recombination suggests that maize genes tend to be recombination hot spots. Timmermans (1996), however, mapped a CO breakpoint to an apparently non-genic region that is low-copy in the maize genome and conserved between the homologous chromosomes. This result suggests that certain non-genic regions may also be recombination active. To differentiate the two hypotheses described above about the relationship between gene density and recombination activity, high resolution mapping of recombination breakpoints relative to both genes and non-genic regions across a large chromosome interval is necessary. A study on how this detailed distribution of meiotic recombination events is regulated via genetic modifiers will help to understand better the mechanisms of meiotic recombination in plant.

***Cis*- and *trans*-modifiers of meiotic recombination in maize.**

Variation in the rates of meiotic recombination in maize has long been observed (STADLER 1926). Recombination frequencies between marker loci are highly variable among a wide range of maize inbreds, wide crosses and maize-teosinte hybrids (WILLIAMS *et al.* 1995). This difference in recombination frequencies can be attributed to the actions of both *trans*- and *cis*-genetic modifiers. *Trans*- and/or *cis*-genetic modifiers may also be responsible for the different distributions of recombination breakpoints among loci. Recombination breakpoints clustered at 5' ends of the *a1* (XU *et al.* 1995) and *b1* (PATTERSON *et al.* 1995) loci and at the 3' end of the *r1* locus (EGGLESTON *et al.* 1995). On the other hand, recombination breakpoints distributed fairly randomly across the *bz1* (DOONER and MARTINEZ-FEREZ 1997) and *wx1* (OKAGAKI and WEIL 1997) loci. Mutations of genes that function directly or indirectly in meiotic recombination pathways could have *trans*-effects on meiotic recombination as is true in yeast (reviewed by PAQUES and HABER 1999). Different genders and genetic backgrounds also affect meiotic recombination in *trans*. In maize, differences of CO frequencies between male and female meiocytes were found in some intervals near heterochromatin structures such as centromeres, knobs, and the B

heterochromatin in a rearranged chromosome from B-A translocation; usually, male meiocytes exhibit higher CO frequency (STADLER 1926; PHILLIPS 1969; NEL 1975; CHANG and KIKUDOME 1974; ROBERTSON 1984). Both the abnormal chromosome 10 (K10) and supernumerary B chromosomes increase recombination frequencies in regions of other chromosomes. These regions are usually close to heterochromatin and/or located in chromosomes with heterozygous chromosome structures (e.g., knob/knobless chromosome pair) (KIKUDOME 1959; CHANG and KIKUDOME 1974). B chromosomes can also change the distribution of COs and increase the frequency of double COs in other chromosomes (HANSON 1969). More recently, region-specific *trans*-modifiers of meiotic recombination have been reported in maize (TIMMERMANS *et al.* 1997). Moreover, a candidate gene that regulates meiotic recombination was identified via genetic and cytogenetic characterizations (JI *et al.* 1999). CO frequencies are generally lower in the mutant (*desynaptic*) than the wild type (JI *et al.* 1999). In addition, the autonomous *MuDR* transposon increases in *trans* the frequency of meiotic COs at the *a1* locus in a stock that carries the non-autonomous *Mu1* transposon insertion at the 5' end of *a1* (YANDEAU *et al.* submitted). This increase is most likely due to the increased level of DSBs induced by the *MuDR* at the *Mu1* insertion site.

Unlike *trans*-modifiers, by definition *cis*-modifiers only affect meiotic recombination in their vicinity. CO frequencies are reduced in some regions between chromosomes that are heterozygous for translocations (ROBERTSON 1967; PHILLIPS 1969). Heterozygosity of the heterochromatic knobs (e.g., a pair of knob/knobless chromosomes) also reduces meiotic recombination in its vicinity (reviewed by RHOADES 1978), i.e., plants that carry K10 and normal chromosome 10 exhibit decreased CO frequencies nearby though the K10 can increase recombination frequencies in other chromosomes (KIKUDOME 1959). Heterochromatic centromeres also suppress meiotic recombination in regions nearby (reviewed by CARLSON 1977). In addition, high-resolution characterization of intragenic recombination revealed that sequence polymorphisms have *cis*-effects on meiotic

recombination. Such sequence polymorphisms include single nucleotide polymorphisms (SNPs) and small or large InDels such as non-autonomous transposon insertions (reviewed by SCHNABLE *et al.* 1998). The high density of small nucleotide heterologies (SNPs and InDels) and large InDels reduce the rate of recombination and change the distribution of recombination breakpoints and the ratio of COs to NCOs at the *bz1* loci (DOONER and MARTINEZ-FEREZ 1997; DOONER 2002). In contrast, a 1.4-kb InDel that resulted from the *Mu1* transposon insertion at the *a1* loci only reduces the rate of recombination but does not change the distribution of recombination breakpoints (XU *et al.* 1995).

References

ALANI, E., R. PADMORE and N. KLECKNER, 1990 Analysis of wild-type and *rad50* mutants of yeast suggests an intimate relationship between meiotic chromosome synapsis and recombination. Cell **61:** 419-436.

ALLERS, T., and M. LICHTEN, 2001 Differential timing and control of noncrossover and crossover recombination during meiosis. Cell **106:** 47-57.

ANDERSON, L. K, G. G. DOYLE, B. BRIGHAM, J. CARTER, K. D. HOOKER, A. LAI, M. RICE and S. M. STACK, 2003 High-resolution crossover maps for each bivalent of *Zea mays* using recombination nodules. Genetics **165:** 849-865.

ARMSTRONG, S. J, A. P. CARYL, G. H. JONES and F. C. FRANKLIN, 2002 Asy1, a protein required for meiotic chromosome synapsis, localizes to axis-associated chromatin in Arabidopsis and Brassica. J. Cell Sci. **115:** 3645-3655.

BAGHERIEH-NAJJAR, M. B., O. M. DE VRIES, J. T. KROON, E. L. WRIGHT, K. M. ELBOROUGH, J. HILLE and P. P. DIJKWEL, 2003 Arabidopsis RecQsim, a plant-specific member of the RecQ helicase family, can suppress the MMS hypersensitivity of the yeast *sgs1* mutant. Plant Mol. Biol. **52:** 273-284.

BAUDAT, F., and A. NICOLAS 1997 Clustering of meiotic double-strand breaks on yeast chromosome III. Proc. Natl. Acad. Sci. USA **94:** 5213-5218.

BERGERAT, A., B. DE MASSY, GADELLE D., P. C. VAROUTAS, A. NICOLAS and P. FORTERRE, 1997 An atypical topoisomerase II from Archaea with implications for meiotic recombination. Nature **386:** 414-417.

BHATT, A. M., C. CANALES and H. G. DICKINSON, 2001 Plant meiosis: the means to 1N. Trends Plant Sci. **6:** 114-121.

BISHOP, D., D. PARK, L. XU and N. KLECKNER, 1992 *DMC1*: a meiosis-specific yeast homolog of *E. coli recA* required for recombination, synaptonemal complex formation, and cell cycle progression. Cell **69:** 439-456.

BORTS, R.H., S. R. CHAMBERS and M. F. ABDULLAH, 2000 The many faces of mismatch repair in meiosis. Mutat. Res. **451:** 129-150.

BRITT, A. B., 1999 Molecular genetics of DNA repair in higher plants. Trends Plant Sci. **4:** 20-25.

BULLARD, S. A., S. KIM, A. M. GALBRAITH and R. E. MALONE, 1996 Double-strand breaks at the *HIS2* recombination hot spot in *Saccharomyces cerevisiae*. Proc. Natl. Acad. Sci. USA 93: 13054-13059.

BUNDOCK, P., and P. HOOYKAAS, 2002 Severe developmental defects, hypersensitivity to DNA-damaging agents, and lengthened telomeres in Arabidopsis MRE11 mutants. Plant Cell **14:** 2451-2462.

CAO, L., E. ALANI and N. KLECKNER, 1990 A pathway for generation and processing of double-strand breaks during meiotic recombination in *S. cerevisiae*. Cell **61:** 1089-1101.

CARLSON, W. R., 1977 The cytogenetics of corn, pp. 225-304 in *Corn and Corn Improvement*, edited by G. F. SPRAQUE. Am. SOC. of Agronomy, Madison, WI.

CARYL, A. P, S. J. ARMSTRONG, G. H. Jones and F. C. Franklin, 2000 A homolo.gue of the yeast *HOP1* gene is inactivated in the Arabidopsis meiotic mutant *asy1*. Chromosoma **109:** 62-71.

CHANG, C. C., and G. Y. KIKUDOME, 1974 The interaction of knobs and B chromosomes of maize in determining the level of recombination. Genetics **77:** 45-54.

CHU, S., J. DERISI, M. EISEN, J. MULHOLLAND, D. BOTSTEIN, P. O. BROWN and I. HERSKOWITZ, 1998 The transcriptional program of sporulation in budding yeast. Science **282:** 699-705.

CHUA, P. R., and G. S. ROEDER, 1997 Tam1, a telomere-associated meiotic protein, functions in chromosome synapsis and crossover interference. Genes Dev. **11:** 1786-1800.

CIVARDI, L., Y. XIA, K. J. EDWARDS, P. S. SCHNABLE and B. J. NIKOLAU, 1994 The relationship between genetic and physical distances in the cloned *a1-sh2* interval of the

Zea mays L. genome. Proc. Natl. Acad. Sci. USA **91:** 8268-8272.

CLYNE, R. K., V. L. KATIS, L. JESSOP, K. R. BENJAMIN, I. HERSKOWITZ, M. LICHTEN and K. NASMYTH, 2003 Polo-like kinase Cdc5 promotes chiasmata formation and cosegregation of sister centromeres at meiosis I. Nat. Cell Biol. **5:** 480-485.

COHEN P. E., and J. W. POLLARD, 2001 Regulation of meiotic recombination and prophase I progression in mammals. Bioessays **23:** 996-1009.

COLLINS, I., and C.S. NEWLON, 1994 Meiosis-specific formation of joint DNA molecules containing sequences from homologous chromosomes. Cell **76:** 65-75.

COPENHAVER, G. P., W. E. BROWNE and D. PREUSS, 1998 Assaying genome-wide recombination and centromere functions with Arabidopsis tetrads. Proc. Natl. Acad. Sci. USA **95:** 247-252.

COUTEAU, F., F. BELZILE, C. HORLOW, O. GRANDJEAN, D. VEZON and M. P. Doutriaux, 1999 Random chromosome segregation without meiotic arrest in both male and female meiocytes of a *dmc1* mutant of Arabidopsis. Plant Cell **11:** 1623-1634.

CULLIGAN, K. M., and J. B. HAYS, 2000 Arabidopsis MutS homologs *AtMSH2*, *AtMSH3*, *AtMSH6*, and a novel *AtMSH7* form three distinct protein heterodimers with different specificities for mismatched DNA. Plant Cell **12:** 991-1002.

DAOUDAL-COTTERELL, S., M. E. GALLEGO and C. I. WHITE, 2002 The plant Rad50-Mre11 protein complex. FEBS Lett. **516:** 164-166.

DE MASSY, B., 2003 Distribution of meiotic recombination sites. Trends Genet. **19:** 514-522.

DOONER, H. K., 2002 Extensive interallelic polymorphisms drive meiotic recombination into a crossover pathway. Plant Cell **14:** 1173-1183.

DOONER, H. K., and I. M. MARTINEZ-FEREZ, 1997 Recombination occurs uniformly within the bronze gene, a meiotic recombination hotspot in the maize genome. Plant Cell **9:** 1633-1646.

DRESSER, M. E., D. J. EWING, M. N. CONRAD, A. M. DOMINGUEZ, R. BARSTEAD, H. JIANG and T. KODADEK, 1997 *DMC1* functions in a *Saccharomyces cerevisiae* meiotic pathway that is largely independent of the *RAD51* pathway. Genetics **147:** 533-544.

EGGLESTON, W. B., M. ALLEMAN and J. L. KERMICLE, 1995 Molecular organization and germinal instability of R-stippled maize. Genetics **141:** 347-360.

FARIS, J. D., K. M. HAEN and B. S. GILL, 2000 Saturation mapping of a gene-rich recombination hot spot region in wheat. Genetics **154:** 823-835.

FRARY, A., G. G. PRESTING and S. D. TANKSLEY, 1996 Molecular mapping of the centromeres of tomato chromosomes 7 and 9. Mol. Gen. Genet. **250:** 295-304.

FU, H., Z. ZHENG and H. K. DOONER, 2002 Recombination rates between adjacent genic and retrotransposon regions in maize vary by 2 orders of magnitude. Proc. Natl. Acad. Sci. USA **99:** 1082-1087.

GALLEGO, M. E., M. JEANNEAU, F. GRANIER, D. BOUCHEZ, N. BECHTOLD and C. I. WHITE, 2001 Disruption of the Arabidopsis *RAD50* gene leads to plant sterility and MMS sensitivity. Plant J. **25:** 31-41.

GERTON, J. L., and J. L. DERISI, 2002 Mnd1p: an evolutionarily conserved protein required for meiotic recombination. Proc. Natl. Acad. Sci. USA **99**: 6895-6900.

GERTON, J. L., J. DERISI, R. SHROFF, M. LICHTEN, P. O. BROWN and T. D. PETES, 2000 Global mapping of meiotic recombination hotspots and coldspots in the yeast *Saccharomyces cerevisiae*. Proc. Natl. Acad. Sci. USA **97:** 11383-11390.

GHERBI, H., M. E. GALLEGO, N. JALUT, J. M. LUCHT, B. HOHN, C. I. WHITE, 2001 Homologous recombination in planta is stimulated in the absence of *Rad50*. EMBO Rep. **2:** 287-291.

GILBERTSON, L. A., and F. W. STAHL, 1996 A test of the double-strand break repair model for meiotic recombination in *Saccharomyces cerevisiae*. Genetics **144:** 27-41.

GILL, K. S., B. S. GILL, T. R. ENDO and E. V. BOYKO, 1996a Identification and high-density mapping of gene-rich regions in chromosome group 5 of wheat. Genetics **143:** 1001-1012.

GILL, K. S., B. S. GILL, T. R. ENDO and T. TAYLOR, 1996b Identification and high-density mapping of gene-rich regions in chromosome group 1 of wheat. Genetics **144:** 1883-1891.

GORBUNOVA, V. V., and A. A. LEVY, 1999 How plants make ends meet: DNA double-strand break repair. Trends Plant Sci. **4:** 263-269.

GOYON C., and M. LICHTEN, 1993 Timing of molecular events in meiosis in *Saccharomyces cerevisiae*: stable heteroduplex DNA is formed late in meiotic prophase. Mol. Cell Biol. **13:** 373-382.

GRELON, M., D. VEZON, G. GENDROT, and G. PELLETIER, 2001 *AtSPO11-1* is necessary for efficient meiotic recombination in plants. EMBO J. **20:** 589-600.

HANSON, G. P., 1969 B-chromosome-stimulated crossing over in maize. Genetics **63:** 601-609.

HARING, S. J., G. R. HALLEY, A. J. JONES and R. E. MALONE, 2003 Properties of natural double-strand-break sites at a recombination hotspot in *Saccharomyces cerevisiae*. Genetics **165:** 101-114.

HARTUNG, F., H. PLCHOVA and H. PUCHTA, 2000 Molecular characterisation of RecQ homologues in *Arabidopsis thaliana*. Nucleic Acids Res. **28:** 4275-4282.

HAUPT, W., T. C. FISCHER, S. WINDERL, P. FRANSZ and R. A. TORRES-RUIZ, 2001 The centromere1 (CEN1) region of *Arabidopsis thaliana*: architecture and functional impact of chromatin. Plant J. **27:** 285-296.

HEYER, W. D., K. T. EHMSEN and J. A. SOLINGER, 2003 Holliday junctions in the eukaryotic nucleus: resolution in sight? Trends Biochem. Sci. **28:** 548-557.

HEPWORTH, S. R., H. FRIESEN and J. SEGALL, 1998 *NDT80* and the meiotic recombination checkpoint regulate expression of middle sporulation-specific genes in *Saccharomyces cerevisiae*. Mol. Cell Biol. **18:** 5750-5761.

HOLLINGSWORTH, N. M., and S. J. BRILL, 2004 The Mus81 solution to resolution: generating meiotic crossovers without Holliday junctions. Genes Dev. **18:** 117-125.

HUNTER N., and N. KLECKNER, 2001 The single-end invasion: an asymmetric intermediate at the double-strand break to double-holliday junction transition of meiotic recombination. Cell **106:** 59-70.

INTERNATIONAL HUMAN GENOME SEQUENCING CONSORTIUM, 2001 Initial sequencing and analysis of the human genome. Nature **409:** 860-921.

JI, Y., D. M. STELLY, M. DE DONATO, M. M. GOODMAN and C. G. WILLIAMS, 1999 A candidate recombination modifier gene for *Zea mays L.* Genetics **151:** 821-830.

KAUPPI, L., A. J. JEFFREYS and S. Keeney, 2004 Where the crossovers are: recombination distributions in mammals. Nat. Rev. Genet. **5:** 413-424.

KEENEY, S., C. N. GIROUX, and N. KLECKNER, 1997 Meiosis-specific DNA double-strand breaks are catalyzed by Spo11, a member of a widely conserved protein family. Cell **88:** 375-384.

KIKUDOME G. Y., 1959 Studies on the phenomenon of preferential segregation in maize. Genetics **44:** 815-831.

KOROL, A. B., I. A. PREYGEL and S. I. PREYGEL, 1994 *Recombination Variability and Evolution*. Chapman and Hall, London.

KUNZEL, G., L. KORZUN and A. MEISTER, 2000 Cytologically integrated physical restriction fragment length polymorphism maps for the barley genome based on translocation breakpoints. Genetics **154:** 397-412.

KUPIEC, M., 2000 Damage-induced recombination in the yeast *Saccharomyces cerevisiae*. Mutat. Res. **451:** 91-105.

LAMBIE, E. J., and G. S. ROEDER, 1988 A yeast centromere acts in *cis* to inhibit meiotic gene conversion of adjacent sequences. Cell **52:** 863-873.

LEONARD, J. M., S. R. BOLLMANN and J. B. HAYS, 2003 Reduction of stability of Arabidopsis genomic and transgenic DNA-repeat sequences (microsatellites) by inactivation of AtMSH2 mismatch-repair function. Plant Physiol. **133:** 328-338.

LICHTEN, M., and A. S. GOLDMAN, 1995 Meiotic recombination hotspots. Annu. Rev. Genet. **29:** 423-444.

LIU, Y., J. Y. MASSON, R. SHAH, P. O'REGAN and S. C. WEST, 2004 RAD51C is required for Holliday junction processing in mammalian cells. Science **303:** 243-246.

LOUIS, E. J., and R. H. BORTS, 2003 Meiotic recombination: too much of a good thing? Curr. Biol. **13:** R953-955.

MURAKAMI, H., V. BORDE, T. SHIBATA, M. LICHTEN and K. OHTA, 2003 Correlation between premeiotic DNA replication and chromatin transition at yeast recombination initiation sites. Nucleic Acids Res. **31:** 4085-4090.

NAG, D. K., and T. D. PETES, 1993 Physical detection of heteroduplexes during meiotic recombination in the yeast *Saccharomyces cerevisiae*. Mol. Cell Biol. **13:** 2324-2331.

NAKAGAWA, T., and H. OGAWA, 1999 The *Saccharomyces cerevisiae MER3* gene, encoding a novel helicase-like protein, is required for crossover control in meiosis. EMBO J. **18:** 5714-5723.

NEL, P. M., 1975 Crossing over and diploid egg formation in the elongate mutant of maize. Genetics **79:** 435-450.

OKAGAKI, R. J., and C. F. WEIL, 1997 Analysis of recombination sites within the maize waxy locus. Genetics **147:** 815-821.

OSAKABE, K., T. YOSHIOKA, H. ICHIKAWA and S. TOKI, 2002 Molecular cloning and characterization of *RAD51*-like genes from *Arabidopsis thaliana*. Plant Mol. Biol. **50:** 71-81.

PADMORE, R., L. CAO and N. KLECKNER, 1991 Temporal comparison of recombination and synaptonemal complex formation during meiosis in *S. cerevisiae*. Cell **66:** 1239-1256.

PAQUES, F., and J. E. HABER, 1999 Multiple pathways of recombination induced by double-strand breaks in *Saccharomyces cerevisiae*. Microbiol. Mol. Biol. Rev. **63:** 349-404.

PATTERSON, G. I., K. M. KUBO, T. SHROYER and V. L. CHANDLER, 1995 Sequences required for paramutation of the maize b gene map to a region containing the promoter and upstream sequences. Genetics **140:** 1389-1406.

PETES, T. D., 2001 Meiotic recombination hot spots and cold spots. Nat Rev Genet. **2:** 360-369.

PHILLIPS, R. L., 1969 Recombination in *Zea Mays L.* II. Cytogenetic studies of recombination in reciprocal crosses. Genetics **61:** 117-127.

PORTER, S. E., M. A. WHITE and T. D. PETES, 1993 Genetic evidence that the meiotic recombination hotspot at the *HIS4* locus of Saccharomyces cerevisiae does not represent a site for a symmetrically processed double-strand break. Genetics **134:** 5-19.

PRADO, F., and A. AGUILERA, 2003 Control of cross-over by single-strand DNA resection. Trends Genet. **19:** 428-431.

PUCHTA, H., and B. HOHN, 1996 From centiMorgans to base pairs: homologous recombination in plants. Trends Genet. **1:** 340-348.

QIN, J., L. L. RICHARDSON, M. JASIN, M. A. HANDEL and N. ARNHEIM, 2004 Mouse strains with an active *H2-Ea* meiotic recombination hot spot exhibit increased levels of *H2-Ea*-specific DNA breaks in testicular germ cells. Mol. Cell. Biol. **24:** 1655-1666.

RHOADES, M. M., 1978 Genetic effects of heterochromatin in maize, pp. 641-671 in *Maize Breeding and Genetics*, edited by D. B. WALDEN. Wiley & Sons, New York.

ROBERTSON, D. S., 1967 Crossing over and chromosomal segregation involving the B9 element of the A-B translocation B-9b in maize. Genetics **55:** 433-449.

ROBERTSON, D. S., 1984 Different frequency in the recovery of crossover products from male and female gametes of plants hypoploid for B-A translocation in maize. Genetics **107:** 117-130.

ROCKMILL, B., J. C. FUNG, S. S. BRANDA and G. S. ROEDER, 2003 The Sgs1 helicase regulates chromosome synapsis and meiotic crossing over. Curr. Biol. **13:** 1954-1962.

SCHMIDT, R. J., K. L. LENEHAN, Z. WEST, C. LISTER, H. THOMPSON, D. BOUCHEZ and C. DEAN, 1995 Physical map and organization of *Arabidopsis thaliana* chromosome 4. Science **20:** 480-483.

SCHNABLE, P. S., A. P. HSIA and B. J. NIKOLAU, 1998 Genetic recombination in plants. Curr. Opin. Plant Biol. **1:** 123-129.

SCHWACHA, A., and N. KLECKNER, 1994 Identification of joint molecules that form frequently between homologs but rarely between sister chromatids during yeast meiosis. Cell **76:** 51-63.

SCHWACHA, A., and N. KLECKNER, 1995 Identification of double Holliday junctions as intermediates in meiotic recombination. Cell **83:** 783-791.

SCHWACHA, A., and N. KLECKNER, 1997 Interhomolog bias during meiotic recombination: meiotic functions promote a highly differentiated interhomolog-only pathway. Cell **90:** 1123-1135.

SCHWARZACHER, T., 2003 Meiosis, recombination and chromosomes: a review of gene isolation and fluorescent in situ hybridization data in plants. J. Exp. Bot. **54:** 11-23.

SHINOHARA, M., K. SAKAI, A. SHINOHARA and D. K. BISHOP, 2003 Crossover interference in *Saccharomyces cerevisiae* requires a TID1/RDH54- and DMC1-dependent pathway. Genetics **163:** 1273-1286.

SMITH, K. N., and A. NICOLAS, 1998 Recombination at work for meiosis. Curr. Opin. Genet. Dev. **8:** 200-211.

STADLER, L. J., 1926 The variability of crossing over in maize. Genetics **11:** 1-37.

SUN, H., D. TRECO, N. P. SCHULTES and J. W. SZOSTAK, 1989 Double-strand breaks at an initiation site for meiotic gene conversion. Nature **338:** 87-90.

SUN, H., D. TRECO and J. W. SZOSTAK, 1991 Extensive 3'-overhanging, single-stranded DNA associated with the meiosis-specific double-strand breaks at the *ARG4* recombination initiation site. Cell **64:** 1155-1161.

SYMINGTON, L. S., 2002 Role of *RAD52* epistasis group genes in homologous recombination and double-strand break repair. Microbiol. Mol. Biol. Rev. 66: 630-670.

SZOSTAK J. W., T. L. ORR-WEAVER, R. J. ROTHSTEIN and F. W. STAHL, 1983 The double-strand-break repair model for recombination. Cell **33:** 25-35.

TANKSLEY, S. D., M. W. GANAL, J. P. PRINCE, M. C. DE VICENTE and M. W. BONIERBALE, *et al.*, 1992 High density molecular linkage maps of the tomato and potato genomes. Genetics **132:** 1141-1160.

THURIAUX, P., 1977 Is recombination confined to structural genes on the eukaryotic genome? Nature **268:** 460-462.

TIMMERMANS, M. C., O. P. DAS, and J. MESSING, 1996 Characterization of a meiotic crossover in maize identified by a restriction fragment length polymorphism-based method. Genetics **143:** 1771-1783.

TIMMERMANS, M. C., O. P. DAS, J. M. BRADEEN and J. MESSING, 1997 Region-specific *cis*- and *trans*-acting factors contribute to genetic variability in meiotic recombination in maize. Genetics **146:** 1101-1113.

VAN DEN BOSCH, M., P. H. LOHMAN and A. PASTINK, 2002 DNA double-strand break repair by homologous recombination. Biol. Chem. **383:** 873-892.

VAN GENT, D. C., J. H. HOEIJMAKERS and R. KANAAR, 2001 Chromosomal stability and the DNA double-stranded break connection. Nat. Rev. Genet. **2:** 196-206.

VERGUNST, A. C., and P. J. HOOYKAAS, 1999 recombination in the plant genome and its application in biotechnology. Critical Reviews in Plant Sciences **18:** 1-31.

WERNER, J. E., T. R. ENDO and B. S. GILL, 1992 Toward a cytogenetically based physical map of the wheat genome. Proc. Natl. Acad. Sci. USA **89:** 11307-1111.

WEST, S. C., 1996 The RuvABC proteins and holliday junction processing in *Escherichia coli*. Journal of Bacteriology **178:** 1237-1241.

WEST, S. C., 1997 Processing of recombination intermediates by the RuvABC proteins. Annu. Rev. Genet. **31:** 213-244.

WILLIAMS, C. G., M. M. GOODMAN and C. W. STUBER, 1995 Comparative recombination distances among *Zea mays L.* inbreds, wide crosses and interspecific hybrids. Genetics **141:** 1573-1581.

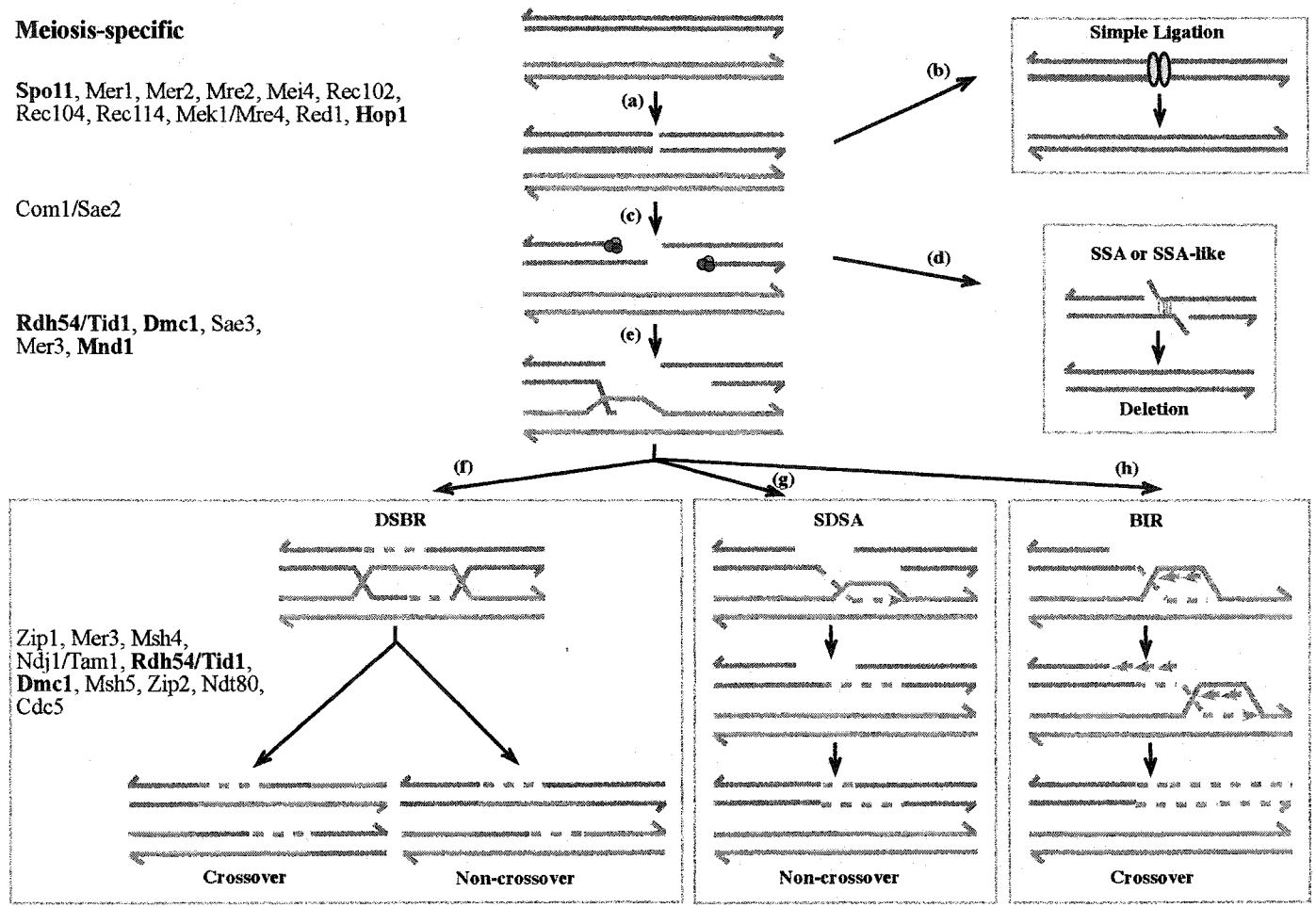
WU, T. C., and M. LICHTEN, 1994 Meiosis-induced double-strand break sites determined by yeast chromatin structure. Science **263:** 515-518.

XU, X., A. P. HSIA, L. ZHANG, B. J. NIKOLAU and P. S. SCHNABLE, 1995 Meiotic recombination break points resolve at high rates at the 5' end of a maize coding sequence. Plant Cell **7:** 2151-2161.

YIN, Y., H. CHEONG, D. FRIEDRICHSEN, Y. ZHAO, J. HU, S. MORA-GARCIA and J. CHORY, 2002 A crucial role for the putative Arabidopsis topoisomerase VI in plant growth and development. Proc. Natl. Acad. Sci. USA **99:** 10191-10196.

Figure 1: Pathways to repair DSBs. (a) DSBs can be generated via processes such as enzymatic reactions (e.g., attack by SPO11 or endonucleases, irradiation, or transposon excision). (b) Rarely, DSBs can be ligated without any change in the sequence. (c) More frequently, resectioning exposes 3' single-stranded overhangs. (d) This intermediate can be repaired via the SSA HR or SSA-like NHEJ pathways, which require long stretches of sequence identity or only several base pairs of identity between the two 3' single-stranded overhangs, respectively. Repair is completed by removal of the non-complementary overhanging 3' ends, DNA synthesis and ligation. Consequently, this pathway will usually generate deletions. (e) Alternatively, the 3' single-stranded overhang can invade a homologous template. Repair can be completed via (f) the DSBR pathway, which results in crossover and/or non-crossover products, (g) the SDSA pathway, which generates only non-crossover products, or (h) the BIR pathway, which generate crossover products, but results in the loss of a portion of the broken chromosome. Proteins involved in meiotic recombination pathways (DSBR and SDSA) are indicated and divided into two groups depending on whether or not they are specific for meiotic recombination. Homologues of proteins in bold have been identified in plants. (Adapted from GORBUNOVA and LEVY 1999)

Meiosis/mitosis	Meiosis-specific
Rad50/Rad50S, **Mre11**/Mre11S, Xrs2	**Spo11**, Mer1, Mer2, Mre2, Mei4, Rec102, Rec104, Rec114, Mek1/Mre4, Red1, **Hop1**
Rad50, **Mre11**, Xrs2?	Com1/Sae2
Rad51, Rad52, **Rad54**, Rad55, **Rad57**, **Rpa1**	**Rdh54/Tid1**, **Dmc1**, Sae3, Mer3, **Mnd1**
Msh2, **Msh3**, **Msh6**, **Pms1**, **Mlh1**	
Mlh1, **Sgs1**	Zip1, Mer3, Msh4, Ndj1/Tam1, **Rdh54/Tid1**, **Dmc1**, Msh5, Zip2, Ndt80, Cdc5



CHAPTER 2. MOLECULAR CHARACTERIZATION OF MEIOTIC RECOMBINATION ACROSS THE 140-KB MULTIGENIC *a1-sh2* INTERVAL OF MAIZE

A paper published in *PNAS*[1]

Hong Yao, Qing Zhou, Jin Li, Heather Smith, Marna Yandeau,

Basil J. Nikolau, and Patrick S. Schnable

Abstract

The 140-kb *a1-sh2* interval of the maize genome contains at least four genes (*a1*, *yz1*, *x1* and *sh2*). Partial sequence analysis of two haplotypes has revealed many single nucleotide polymorphisms and InDel polymorphisms, including several large structural polymorphisms. The physical positions of 101 meiotic recombination breakpoints are not distributed uniformly across the interval and are instead concentrated within three recombination hot spots. Two of these recombination hot spots are genic (*a1* and *yz1*) and one is apparently non-genic. The *x1* gene is not a recombination hot spot. Thus, these results suggest that not all hot spots are genes and indicate that not all genes are hot spots. Two of the 101 recombination events arose by means of either noncrossover events involving conversion tract lengths of at least 17 kb or double-crossover events. Only one recombination breakpoint mapped to the ≈80-kb distal portion of the *a1-sh2* interval that contains large amounts of repetitive DNA including retrotransposons; in this region the ratio of genetic to physical distance is less than 0.5% of the genome's average. These results establish that the retrotransposon fraction of the maize genome is relatively inert recombinationally.

Introduction

Homologous meiotic recombination recombines physically linked genetic material via reciprocal crossovers (COs) and unilateral NCOs. According to the canonical double-

---

[1] Reprinted with permission of *PNAS*, 2002, 99(9), 6157-6162.

strand break (DSB) repair model (1, 2) a meiotic recombination event of either type is initiated by a DSB and depending on how the recombination intermediate, a double Holliday junction (DHJ), is resolved, a CO or NCO results. Recently, a modified DSB repair model has been proposed in which an early commitment is made to enter either the CO or NCO pathway (3).

Bacterial, fungal, plant and mammalian genomes all exhibit recombination "hot spots" and "cold spots", where recombination rates per kb are much higher or lower than the genome average (4–6). Even though the sizes of the genomes of diverse eukaryotic organisms are quite different, the lengths of their genetic maps are fairly constant. Based on this observation, and the assumption (now being confirmed by genome sequencing projects) that these genomes contain similar numbers of genes, it was hypothesized that recombination occurs primarily in genes (7). Several observations are consistent with this hypothesis: (*i*) all recombination hot spots identified to date in the approximate 2,500-Mb and 5,289-cM (centimorgan) (Georgia Davis, personal communication) maize genome are genes (6), even though the bulk of this genome consists of repetitive DNA such as retrotransposons (8); (*ii*) gene-rich chromosomal regions of wheat (9–11) and barley (12) are more recombinationally active than gene-poor regions; and (*iii*) in *Arabidopsis* (13) and tomato (14, 15) recombination is suppressed in chromosomal regions near the gene-poor centromeres (16, 17).

Two alternative hypotheses have been proposed to explain the correlation between recombination rates and gene density (6). One is that genes *per se* are recombinationally hyperactive such that chromosomal regions with high gene densities are recombinationally hyperactive. The alternative hypothesis is that genes *per se* do not exhibit recombinational

hyperactivity, but genes tend to cluster in recombinationally hyperactive chromosomal regions. Analyses of recombination within single genes can not distinguish between these hypotheses. Instead, it is necessary to characterize the distribution of recombination events across an interval that contains a mixture of genic and non-genic regions.

We tested these hypotheses by characterizing the distribution of recombination events across the 140-kb multigenic *a1-sh2* interval of the maize genome (18). We address the questions of whether all genes in this interval are recombination hot spots and whether all hot spots in this interval are genes. The *a1-sh2* interval (GenBank accession nos. AF072704, AF347696, AF434192, and AF434193) is an ideal model for such studies because recombination events across this interval can be readily selected via phenotypic screens, and this interval contains a mixture of genic and nongenic regions.

Materials and Methods

Maize Genetic Stocks. The *A1-LC* allele was derived from the inbred Line C and conditions a colored kernel phenotype. The *a1::rdt* and *a1-mum2* alleles contain transposons that disrupt the function of the *a1* gene (19–21) and condition a colorless phenotype in the stocks used in this study. Kernels that carry functional *Sh2* alleles are round whereas those homozygous for *sh2* alleles are shrunken (22).

Restriction Fragment Length Polymorphism (RFLP) Markers. Seven cosmid subclones (in SuperCos1, Stratagene) of yeast artificial chromosome ASH-2 (18) constitute a contig spanning almost the entire 140-kb *a1-sh2* interval (Fig. 1). The 9-10a5-800 and 2-32a2-1000 RFLP markers were isolated as plasmid subclones (in pBSK, Stratagene) from these cosmids.

The single-copy *9-10a5* locus includes a 3' portion of the *yz1* gene and is detected with the 0.8-kb *Hind*III-*Sac*IA insert released from pUCA5. The 1.0-kb fragment that detects the *2-32a2* locus was PCR amplified from clone pUCA2 using vector-derived primers. This probe detects several loci in the maize genome. One of them, a 5.5-kb *Bgl*II fragment, maps genetically to the *9-10a5/yz1-sh2* interval and is present in the *a1-sh2* interval from the *A1-LC Sh2* but not the *a1::rdt sh2* haplotype. The RFLP probes that detect the *php10080*, *a1* and *sh2* loci have been described previously (23–25). Computational analysis of the sequences of the rice and sorghum *a1-sh2* intervals revealed a predicted gene (26, 27). Its maize homolog, *x1*, was identified and physically positioned in the *a1-sh2* interval (see additional *Methods*, which are published as supporting information on the PNAS web site, www.pnas.org). The structure of the single-copy *x1* gene was determined by sequencing clones that contain *x1*-cDNA and genomic DNA (unpublished work). A 400-bp *x1* gene-specific probe was obtained via PCR amplification of a 1.4-kb cDNA clone (X-V3) using primers that anneal to its last exon.

Sequence of the *a1-sh2* Interval. Portions of the *a1-sh2* interval from the two haplotypes (*A1-LC Sh2* and *a1::rdt sh2*) from which recombinants were isolated were sequenced. The *a1* to *yz1* interval from the *a1::rdt sh2* haplotype is 12,558 bp (GenBank accession no. AF072704). A 21,230-bp sequence that spans the *a1* and *yz1* interval was assembled from sequences of clones derived from the *A1-LC Sh2* haplotype [GenBank accession nos. X05068 (24), AF363390, and AF363391] and two closely related haplotypes [*a1-mum2 Sh2* (GenBank accession no. AF347696) and *A1-LH82 Sh2* (GenBank accession no. AF434192)]. The structures of the *A1-LC Sh2*, *a1-mum2 Sh2*, and *A1-LH82 Sh2* haplotypes are identical or

nearly identical in the region between the *a1* and *yz1* loci (see additional *Methods*). Hence, the 21,230-bp sequence has been designated the "*A1-LC Sh2*" haplotype.

The sequence of the *x1* allele (GenBank accession no. AF434193) from the inbred LH82 was obtained by sequencing portions of cosmid Cos2-32. Partial sequences of the *x1* alleles from the Line C and *a1::rdt sh2* stocks were obtained by sequencing PCR products amplified from the corresponding genomic DNAs (GenBank accession nos. AF434194 and AF434195).

Oligonucleotide Design. PCR and sequencing primers were designed based on the sequences of the "*A1-LC Sh2*" and *a1::rdt sh2* haplotypes. Allele-specific primers were designed according to InDel polymorphisms (IDPs) between these haplotypes. Nonspecific universal primers also were designed that anneal to both haplotypes. In all instances primers designed based on the "*A1-LC Sh2*" sequence behaved as expected when used to amplify genomic DNA from Line C. Primer sequences are in the additional *Methods*.

Results

Sequence Analyses of the 140-kb *a1-sh2* Interval. Analyses of large portions of the cosmid contig revealed the presence of two new genes, *yz1* and *x1* (GenBank accession nos. AF434192 and AF434193, unpublished work). The *A1-LC Sh2* and *a1::rdt sh2* haplotypes exhibit both small sequence heterologies and large structural polymorphisms (Fig. 2). The 1.1-kb TD1 and TD2 sequences comprise a tandem duplication that is present in the *A1-LC Sh2* haplotype. TD2 contains a 0.6-kb novel MITE (miniature inverted repeat transposable element) (28) termed *Gnat1*. *Gnat1*-like sequences occur about 2,000 times in the maize

genome (Y. Fu and P.S. Schnable, unpublished observation). The *a1::rdt sh2* haplotype contains only TD1. Two retrotransposons, *Ozymandias* and *Machiavelli*, are present in the *A1-LC Sh2* but not the *a1::rdt sh2* haplotype. *Ozymandias* is incomplete and contains only its two long terminal repeats and the primer-annealing site. *Machiavelli* is an apparently intact 6.2-kb *Ty1/copia*-like retrotransposon.

Isolation of Recombination Events. Meiotic recombination events that resolved within the 140-kb *a1-sh2* interval were isolated from the test cross: *A1-LC Sh2/a1::rdt sh2* x *a1::rdt sh2/a1::rdt sh2* based on their non-parental recombinant phenotypes. From a population of 249,000 progeny, 78 colored shrunken and 165 colorless round kernels were isolated, which presumably carried recombinant chromosomes, designated *A1* sh2* and *a1* Sh2*. The genotypes of about half of these exceptional progeny were tested by backcrosses to the *a1::rdt sh2* stock. The fractions of the two classes of progeny that were verified to carry recombinant chromosomes were used to determine the numbers of actual recombinants among the progeny with nonparental phenotypes (Table 1). The genetic distance between *a1* and *sh2* is 0.070 ± 0.005 cM (Table 2, which is published as supporting information on the PNAS web site). Plants homozygous for the recombinant *A1* sh2* and *a1* Sh2* chromosomes were generated via self-pollinations. In total, 101 recombinants were recovered for analysis.

Mapping Recombination Resolution Sites. The resolution sites associated with each recombinant haplotype were mapped relative to molecular markers. Initially, the 101 recombinants were subjected to RFLP analysis using the markers php10080, a1-4300, 9-

10a5/yz-800, x1-400, 2-32a2-1000, and sh2-1000. Such analyses revealed 10 different DNA hybridization patterns (Table 3, which is published as supporting information on the PNAS web site) that represent five classes of CO events (Fig. 2, classes 1–5) and one class of NCO or double crossover (DCO) events (class 6). The class 1 events resulted from reciprocal COs between the *rdt* transposon insertion site in *a1* and E(2) in *Ozymandias* (Fig. 2); the class 2 events resulted from COs between E(2) and E(5) in *Machivalli*; COs between E(5) and B(2) near *x1* resulted in the class 3 events; COs between B(2) and a polymorphic *Bgl*II site in the vicinity of *2-32a2* gave rise to the class 4 events; and COs that resolved between the polymorphic *Bgl*II site near *2-32a2* and a polymorphic *Eco*RI site in the vicinity of *sh2* generated the class 5 event. The class 6 events arose by NCOs or DCOs in which part of the sequence in the *a1::rdt sh2* haplotype was replaced with that from the *A1-LC Sh2* haplotype. The proximal breakpoints associated with these class 6 events mapped between *php10080* and E(1) and the distal breakpoints in *yz1*.

Mapping Recombination Breakpoints to High Resolution. The physical positions of breakpoints associated with recombinants in each of the hybridization classes were more precisely mapped by using IDPs and single nucleotide polymorphisms between the two parental haplotypes.

PCR amplifications using the *a1::rdt sh2*-specific primers QZ1001 and QZ3470 (Fig. 2) coupled with primers that amplify both haplotypes mapped 17 of the 19 class 1 recombination breakpoints to the 1.7-kb Interval II defined by the *rdt* transposon and the QZ1001 annealing site. The remaining two class 1 breakpoints mapped to the 1.8-kb interval III between the QZ1001 annealing site and TD2. Similarly, *A1-LC Sh2*-specific primers

(QZ684, YZ4725, ZH1384 and HYx6488L) and an *a1::rdt sh2*-specific primer (QZ3470), were used to map the recombination breakpoints in the other classes (Fig. 2). Thirty-four class 2 recombination breakpoints mapped to the 2.2-kb interval VI (designated the interloop region, IR) flanked by *Ozymandias* and *Machiavelli* retrotransposons. Thirty-eight of the 39 class 3 recombination breakpoints and the distal breakpoints of the two class 6 events resolved within the 3.4-kb interval VIII that includes *yz1*. The remaining class 3 breakpoint mapped to an approximately 35-kb interval between the ZH1384 and HYx6488L annealing sites (interval IX). All six class 4 breakpoints mapped in the 6.2-kb region containing the *x1* gene (interval X). The single class 5 recombinant mapped to the ≈66-kb interval XIII between *2-32a2* and *sh2*.

Interval II. Seventeen recombination breakpoints mapped to the 1.7-kb interval II that is composed of the 5' portion of the *a1* gene. The ratio between interval II's genetic and physical distances is 6.9 cM/Mb, which is three times higher than the genome's average (2.1 cM/Mb). Therefore, as reported previously (18, 29), the *a1* gene is a recombination hot spot. The 17 recombination breakpoints in interval II were further mapped relative to DNA sequence polymorphisms by cleaved amplified polymorphic sequences and sequence analyses. Interval II was PCR-amplified from plants homozygous for individual recombinant haplotypes (Fig. 3A). The resulting PCR products were subjected to *Pst*I digestion as described by Xu *et al.* (23) and sequenced. These analyses established that 16 recombinant haplotypes had breakpoints distal to the diagnostic *Pst*I site; only one breakpoint mapped proximal to this *Pst*I site (Fig. 3A). Analysis of the PCR product sequences established the physical position of each recombination breakpoint relative to DNA polymorphisms that exist between the two parental haplotypes. As shown in Fig. 3A, 12 of the 17 recombination

breakpoints mapped to the previously identified 377-bp hot spot within the *a1* gene (23). Only four of the 17 breakpoints mapped between the hot spot and the QZ1001 annealing site. The ratio of genetic to physical distances in this 377-bp hot spot is 22 cM/Mb. Hence, this hot spot experiences approximately 10-fold more recombination per unit physical length than the genome's average.

Interval VI. Thirty-four recombinant haplotypes contain breakpoints in the 2.2-kb interval VI (the IR). The breakpoints associated with these recombinant haplotypes were mapped relatively to four polymorphic restriction enzyme recognition sites (*Eco*RV, *Alu*I, *Dde*I, and *Ssp*I) (Fig. 3B) between the parental haplotypes by cleaved amplified polymorphic sequence analyses. In addition, portions of the PCR-amplified IR from these 34 recombinants were sequenced (Fig. 3B). All but one of the 34 recombination breakpoints mapped to the 1.4-kb distal portion of the IR that is flanked by the a1rdt2912 and a1rdt1541 annealing sites (Fig. 3B). This 1.4-kb segment is single-copy in the maize genome (data not shown) and exhibits fewer sequence polymorphisms between the two parental haplotypes than the remainder of the 0.8-kb segment of the IR that is repetitive in the maize genome (Fig. 3B and data not shown).

Several computational approaches were used to test whether the IR contains a gene. BLAST analyses (BLASTN, BLASTX, and TBLASTX) of the IR sequence against GenBank revealed that IR-related sequences are absent from the rice and sorghum *a1-sh2* intervals; the 1.4-kb single-copy distal portion of the IR does not exhibit significant sequence similarity to any Genbank accessions. None of four gene prediction algorithms (FGENESH, GeneMark.hmm, GENSCAN and GlimmerR) predict any genes in the single-copy portion of the IR. Reverse transcription-PCR experiments were performed to test whether the IR is

transcribed. Primers were designed in regions of the IR that contain potential ORFs and are flanked by predicted splice sites (Fig. 3B). No expression was detected in seedling, shoot, adult leaf, tassel, husk or ear tissue. Thus, there is no evidence that the IR is genic. The ratio of genetic to physical distances in the entire IR is 11 cM/Mb. This is approximately five times higher than the genome's average. The 1.4-kb single-copy portion of the IR has a value of 17 cM/Mb, approximately eight times higher than the genome's average. This result establishes that this portion of the apparently non-genic IR is a recombination hot spot.

Interval VIII. The breakpoints associated with the 40 recombination events in interval VIII (the *yz1* gene) flanked by primer sites YZ4725 and ZH1384 were mapped to higher resolution by using three additional pairs of primers (ZH1748/ZH2617, HYyz2222L/ZH1748, and IDPyzrdt/ZH2587) (Fig. 3C). HYyz2222L is a *A1-LC Sh2* haplotype-specific primer whereas IDPyzrdt is specific to the *a1::rdt sh2* haplotype. PCR amplification with ZH1748 and ZH2617 detects an IDP (InDel 2492) that is located in intron 3. These PCR analyses revealed that 16 breakpoints map to the 0.84-kb region defined by the annealing sites of ZH1384 and HYyz2222L and containing the first two exons. Another 19 recombinant breakpoints resolved in the 1.1-kb region flanked by Indel 2492 and the IDPyzrdt annealing site and containing exon 4 and 5. The remaining five recombination breakpoints map to the 1.2-kb region that includes the last two exons. The ratio of genetic to physical distances in *yz1* is 8.2 cM/Mb, a value approximately four times higher than the genome's average. Hence, like all other maize genes studied to date, the *yz1* gene is a recombination hot spot.

Interval X. Six breakpoints resolved within the 6.2-kb Interval X that contains the *x1* gene. The polymorphic primer HYx6488L coupled with the universal primer H9-forward

(Fig. 3D) amplifies genomic DNA from Line C but not from the *a1::rdt sh2* stock. Because the resulting PCR product cosegregates with the *a1-sh2* interval, it was used as a marker to map the recombination resolution sites within interval X. PCR amplification using another *A1-LC Sh2* haplotype-specific primer, XL6, coupled with the universal primer XL3, revealed that all six recombination breakpoints mapped to the 3' end of *x1*. The ratio of genetic to physical distances in the *x1* gene is 0.67 cM/Mb, a value that is much lower than all other maize genes characterized to date and the genome average (2.1 cM/Mb). Hence, unlike all other maize genes studied to date, the *x1* gene is not a recombination hot spot.

Intervals XI-XIV. Only one breakpoint occurred in the 80-kb *x1-sh2* interval. Hence, this region is nearly recombinationally inert. Indeed, the ratio of genetic to physical distances within this region (≈0.0087 cM/Mb) is less than 0.5% of the genome's average.

Discussion

The Retrotransposon Fraction of the Maize Genome is Recombinationally Inert. The positions of recombination resolution sites are not randomly distributed across the 140-kb *a1-sh2* interval. Only one of the 101 recombination events resolved within the ≈80-kb *x1-sh2* interval. The ratio of genetic to physical distances in this half of the *a1-sh2* interval (0.0087 cM/Mb) is less than 0.5% that the genome's average. Based on hybridization data, this subinterval contains large amounts of repetitive DNA (data not shown), some of which is derived from retrotransposons (GenBank accession nos. AF464766 to AF464773). Hence, this result is consistent with the view that the retrotransposon fraction of the maize genome is not recombinationally active.

In contrast, all but eight of 101 recombination breakpoints resolved within the 21-kb *a1-yz1* interval. This interval of the *A1-LC Sh2* and *a1::rdt sh2* haplotypes exhibits three large structural polymorphisms that arose by tandem duplication events and/or transposon/retrotransposon insertions. None of the 93 recombination breakpoints that mapped in the *a1-yz1* interval resolved within these three structural polymorphisms. This finding indicates that at least when hemizygous, retrotransposons can be recombinationally inert. Given that a large fraction of the maize genome is composed of retrotransposons and the highly polymorphic nature of this genome, this finding at least partially explains why recombination events generally cluster within genes.

Identification of a Recombination Hot Spot that is Probably Not a Gene and a Gene that is Not a Hot Spot. Within the 21-kb *a1-yz1* interval recombination resolution sites clustered into three recombination hot spots. No recombination hot spots are located in the remaining 120 kb of the *a1-sh2* interval (i.e., *yz1-sh2*). Thus, these results establish that within the 140-kb multigenic *a1-sh2* interval recombination hot spots cluster in a region (*a1-yz1*) that is larger than a single gene. Two of the recombination hot spots are genic (*a1* and *yz1*) and one is apparently non-genic (the IR). Although Timmermans *et al.* (30) isolated a single recombination event in an apparently non-genic region, it had not previously been established that non-genic regions of a plant genome can serve as recombination hot spots. These data suggest that the hypothesis that all plant recombination hot spots are genic is not correct.

The 6.2-kb *x1* gene exhibits a ratio of genetic to physical distance (0.67 cM/Mb) that is lower than any other characterized maize gene. Even within the most recombinationally

active 3' portion of *x1*, this ratio is only 2.6 cM/Mb, which is approximately equal to the genome's average (2.1cM/Mb). Hence, the *x1* gene can not be considered a recombination hot spot. Although inversions can inhibit recombination, our mapping and sequence data are not consistent with the presence of an inversion in *x1*. Hence, the hypothesis that all genes are recombination hot spots can be rejected.

Some of the unique genic (*a1* and *yz1*) and apparently non-genic (IR) sequences in the *a1-sh2* interval are recombination hot spots. In contrast, the unique genic sequence *x1* is not a recombination hot spot. This suggests that uniqueness within the genome is not a sufficient condition for high recombinational activity. Instead, it is likely that the recombinational activity of a sequence depends in part on its chromatin structure that in turn can be influenced by its regional environment. For example, hot spots for DSB initiation in *S. cerevisiae* generally have open chromatin structure (31, 32). The finding that the maize recombination hot spot *bz1* resides in a 32-kb gene-rich region without retrotransposons (33) is consistent with this view. The unique sequences in the *a1-yz1* interval that host recombination hot spots may contribute synergistically to promote a chromatin structure that is accessible to the recombination machinery. In contrast, the single-copy *x1* locus appears to be isolated from other unique regions, which may reduce its ability to form an open chromatin structure, thereby reducing its accessibility to the recombination machinery. However, the recombinational activity in *x1* is still more than 30 times higher than its flanking regions (0.67 cM/Mb *vs.* 0.02 and 0.0087 cM/Mb). Similarly, the *a1*, *yz1* and IR hotspots also exhibit substantially more recombinational activity than their flanking regions. Hence, these data suggest that regional chromatin structure is not sufficient to create recombination hot or cold spots.

Distribution of Recombination Breakpoints within *a1*, *yz1* and the IR. The distribution of recombination breakpoints within recombinationally active maize genes varies. Breakpoints are distributed fairly uniformly in the *bz1* and *wx1* loci (34, 35). In contrast, breakpoints cluster at the 5' ends of the *a1* (ref. 23 and this study) and *b1* loci (36), and the 3' end of the *r1* locus (37). The two new hot spots defined in this study (IR and *yz1* locus) also exhibit different patterns of breakpoint distribution. Within the IR, almost all breakpoints clustered at the distal portion, whereas breakpoints were distributed relatively uniformly across the *yz1* locus, although its 3' end is somewhat less recombinationally active than the remainder of the gene.

The variation in the distribution of recombination breakpoints within these hot spots may be caused by *cis*-acting modifiers that regulate the resolution of recombination intermediates. For example, sequence heterologies have a major effect on recombination in *Saccharomyces* (38) and other fungi (39). Fewer recombination events resolve in those regions of the *bz1* locus that exhibit high densities of sequence heterologies (34). Consistent with these observations, recombination events preferentially resolved in regions with the highest levels of sequence identity within the *a1* locus and the IR (Fig. 3A and B). However, within the 1.4-kb distal portion of the IR, the density of recombination resolution sites is not correlated with the density of DNA sequence heterologies. In addition, even though the *x1* alleles from the two parental haplotypes do not contain any sequence polymorphisms in the region between exons 2 and 6, no recombination events were observed in that portion of the *x1* gene (Fig. 3D). Hence, although a high level of sequence identity may contribute to the recombinational activity of a sequence, it is not sufficient to define a recombination hot spot.

Effects of Transposon Insertions on Recombination. Rates of intragenic recombination are suppressed in the vicinity of *Ds* and *Mu1* insertions (23, 34, 40). In addition, *Ds* insertions are thought to alter the distribution of recombination breakpoints in the otherwise uniformly recombinogenic *bz1* locus to create allele-specific hot and cold spots (34). In contrast, a preliminary analysis did not provide any evidence that a *Mu1* insertion in the *a1* gene alters the distribution of recombination event (23).

In this previous study, the positions of 15 recombination events isolated from the *a1-mum2/a1::rdt* heterozygote were physically mapped within the 1.2-kb interval of the *a1* gene that is defined by the *Mu1* and *rdt* transposon insertions. All but one of these recombination events resolved within a 377-bp recombination hot spot. Xu *et al.* (23) compared this distribution of recombination events to those isolated from a directly comparable heterozygote that does not contain the *Mu1* insertion in the *a1* gene (*A1-LC/a1::rdt*). This comparison is appropriate because, other than the *Mu1* insertion, the *A1-LC* and *a1-mum2* alleles have identical sequences (GenBank accession nos. X05068, AF363390, AF363391, and AF347696). All four of the recombination events isolated from an *A1-LC/a1::rdt* heterozygote resolved within the 377-bp hot spot. In the current study the positions of an additional 10 intragenic recombination events isolated from the *A1-LC/a1::rdt* heterozygote and that physically mapped within the 1.2-kb region studied by Xu *et al.* (23) were determined. All but two of these recombination events resolved within the 377-bp hot spot that experienced 10-fold more recombination than the genome's average.

Dooner and Martinez-Ferez (34) have suggested that large hemizygous insertions can suppress recombination in nearby regions and thereby create recombination hot spots.

Indeed, they have observed that within the *bz1* locus the lowest ratio of genetic to physical distances occurred in the "interval defined by the insertion and the closest point mutation." Consistent with this observation, none of the 15 recombination breakpoints isolated from *a1-mum2/a1::rdt* by Xu *et al.* (23) resolved within the interval defined by the *Mu1* insertion at –97 in *a1-mum2* and the closest polymorphism. However, at most only one of 17 recombination breakpoints isolated from a *A1-LC/a1::rdt* heterozygote in the current study resolved within this interval. Because the *A1-LC* allele does not contain a *Mu1* insertion, the *Mu1* insertion in *a1-mum2* can not be responsible for the lack of recombination resolution events in the interval defined by positions –97 and +16. In summary, the relative distributions of recombination resolution sites in the 377-bp hot spot defined by Xu *et al.* (23) and the interval defined by positions –97 and +16 are not affected by the presence or absence of the *Mu1* insertion. Hence, these data provide strong support for the view that unlike the *Ds* insertions in *bz1* the *Mu1* insertion in *a1-mum2* does not affect the distribution of recombination resolution sites and is therefore not responsible for the recombination hot spot reported by Xu *et al.* (23).

Implication from Isolation of Two NCO or DCO Events. NCOs unilaterally transfer genetic data from one chromatid to another. Hence, they differ from COs in that the latter event, but not the former, results in the exchange of flanking markers. In plants it is not usually possible to distinguish between NCOs and DCOs. The rate of closely linked DCOs is, however, expected to be low because in the absence of interference the probability of a DCO is the product of the probabilities of two single CO events. The two Class 6 recombination events observed in this study (Fig. 2) could have arisen via either DCOs or

NCOs. Because the genetic distance between *php10080* and *sh2* is 2 cM it would be expected to observe two COs between *a1* and *php10080* among 101 individuals. However, the 101 individuals analyzed in this study each carried a recombination breakpoint between *a1* and *sh2*. Hence, only if the rate of interference in this chromosomal region is very low would the two Class 6 recombination events be likely to have arisen via DCOs. On the other hand, if the Class 6 events represent NCOs, then they involve very long conversion tracts.

Conversion tracts of NCOs in *Drosophila* are relatively short and continuous (41), with a mean length of 350 bp. In contrast, long and interrupted conversion tracts (up to 5.9 kb) have been observed in *Neurospora* (42). Two apparent conversion tracts in the maize *a1* gene are in excess of 590 and 787 bp (23). Another two apparent conversion tracts at the maize *bz1* locus are between 965 and 1165 bp and between 1.1 and 1.5 kb (34). If the class 6 events in the current study are NCOs, they are associated with much longer conversion tracts. One end point of each putative conversion tract is between *php10080* and *a1* and the other is within *yz1*. Hence, if the class 6 events are NCOs, they involve conversion tracts of at least 17 kb.

Acknowledgments

We thank Drs. Curt Hannah for the *sh2* clones, Steve Dellaporta for the W22 cDNA library, and Monica Frey and Alfons Gierl for the C131A cDNA library. Lei Zhang and Dr. Yiji Xia constructed the cosmid contig, Dr. Lisa M. Weaver sequenced clones from interval XIII, and Yuan Zhang constructed and sequenced pCS-YZ. This research was supported by competitive grants from the United States Department of Agriculture-National Research Initiative Program (award numbers: 9701407 and9901579, to P.S.S and B.J.N., and Award

0101869 to P.S.S). This is Journal Paper No J-19604 of the Iowa Agriculture and Home Economics Experiment Station, Project Nos. 3125, 3334, 3485, and 6502, supported by Hatch Act and State of Iowa funds.

Figure Legends

Fig. 1. Physical and genetic characterization of the *a1-sh2* interval. (A) Genetic organization of the *a1-sh2* interval on chromosome 3L. (B) A restriction map of the 140-kb *a1-sh2* interval includes rare cutting restriction enzyme sites (labeled vertical bars; N=*Not*I; P=*Pac*I; A=*Asc*I; S=*Sfi*I) and *Eco*RI sites (unlabelled short vertical bars). The *a1* and *sh2* genes are not drawn to scale. The sizes of the individual cosmids that comprise a contig of the *a1-sh2* interval are shown (the proximal end of this contig is located within TD2, Fig. 2). Unlabeled short vertical bars on the cosmids represent the corresponding rare cutting restriction enzyme sites indicated in the restriction map. Shaded horizontal bars (not drawn to scale) represent RFLP markers.

Fig. 2. The distribution of breakpoints associated with the 101 recombinants across the 140-kb *a1-sh2* interval. The positions of six RFLP markers, *php10080*, *a1*, *9-10a5*, *x1*, *2-32a2*, and *sh2* are indicated relative to the proximal and distal ends of chromosome 3. Intervals I-XIII are defined by the positions of RFLPs and IDPs. Interval IV consists of 1.7 kb in the *A1-LC Sh2* haplotype and 0.4 kb in the *a1::rdt sh2* haplotype. Interval V is approximately 80 bp in both haplotypes. The numbers of recombinants that map to each individual interval and the resulting ratios of genetic to physical distance (cM/Mb) are shown. Figure is not drawn to scale. E = *Eco*RI. B = *Bgl*II.

Fig. 3. High-resolution mapping of recombination breakpoints within (A) the *a1* gene (interval II), (B) the IR (interval VI), (C) the *yz1* gene (interval VIII), and (D) the *x1* gene (interval X). The *A1-LC Sh2* haplotype (gray boxes) is positioned above the *a1::rdt sh2* haplotype (dotted boxes). Vertical bars and parentheses represent DNA sequence polymorphisms between the two haplotypes; IDPs indicated by parentheses were used to design allele-specific primers. The widths of vertical bars are proportional to the numbers of bases in a polymorphism. Polymorphisms in the region flanked by primers a1rdt3273 and a1rdt1541 (panel B) were confirmed to exist in the *A1-LC Sh2* haplotype by sequencing the 34 recombinant alleles that mapped to the IR. The region from the *A1-LC Sh2* and *a1::rdt sh2* haplotypes that are flanked by primers XL1 and x502 (panel D) were sequenced. Primers used in PCR and sequencing are indicated by horizontal arrows. Universal primers are indicated in italics. Primers used for RT-PCR are underlined. Allele-specific primers are positioned close to the haplotypes they amplify. The numbers of recombination breakpoints that resolved within each interval and the resulting ratios of genetic to physical distance (cM/Mb) are shown. The triangle represents the *rdt* transposon insertion in the *a1::rdt* allele. Restriction enzyme sites used in cleaved amplified polymorphic sequence analyses are indicated.

References

1. 1Szostak, J. W., Orr-Weaver T. L., Rothstein, R. J., Stahl, F. W. (1983) *Cell* **33,** 25-35.
2. Sun, H., Treco, D., & Szostak, J. W. (1991) *Cell* **64,** 1155-1161.
3. Allers, T. & Lichten, M. (2001) *Cell* **106,** 47-57.
4. Lichten, M. & Goldman, A. S. H. (1995) *Annu. Rev. Genet.* **29,** 423-444.

5. Puchta, H. & Hohn, B. (1996) *Tren. Genet.* **1,** 340-348.

6. Schnable, P. S., Hsia, A.-P., Nikolau, B. J. (1998) *Curr. Opin. Plant Bio.* **1,** 123-129.

7. Thuriaux P. (1977) *Nature.* **268,** 460-462.

8. SanMiguel, P., Tikhonov, A., Jin, Y.-K., Melake-Berhan A., Springer P. S., Edwards K. J., Avramova Z. & Bennetzen J. L. (1996) *Science* **274,** 765-768.

9. Gill, K. S., Gill, B. S., Endo, T. R. & Boyko E. V. (1996) *Genetics* **143,** 1001-1012.

10. Gill, K. S., Gill, B. S., Endo, T. R. & Taylor, T. (1996) *Genetics* **144,** 1883-1891.

11. Faris, J. D., Haen, K. M. & Gill, B. S. (2000) *Genetics* **154,** 823-835.

12. Künzel, G., Korzun, L. & Meister, A. (2000) *Genetics* **154,** 397-412.

13. Copenhaver, G. P., Browne, W. E. & Preuss, D. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 247-252.

14. Tanksley, S. D., Ganal, M. W., Prince, J. P., de Vicente, M. C., Bonierbale, M. W., Broun, P., Fulton, T. M., Giovannoni, J. J., Grandillo, S., Martin, G. B., *et al.* (1992) *Genetics* **132,** 1141-1160.

15. Frary, A., Presting, G. G., Tanksley, S. D. (1996) *Mol. Gen. Genet.* **250,** 295-304.

16. Moore, G. (2000) *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **51,** 195-222.

17. The Arabidopsis Genome Initiative (2000) *Nature* **408,** 796-815.

18. Civardi, L., Xia, Y. J., Edwards, K. J., Schnable, P. S. & Nikolau, B. J. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 8268-8272.

19. O'Reilly, C., Shepherd, N. C., Pereira, A., Schwarz-Sommer, Z., Bertam, I., Robertson, D. S., Peterson, P. A. & Saedler, H. (1985) *EMBO J.* **4,** 877-882.

20. Shepherd, N. S., Sheridan, W. F., Matters, M. G. & Deno, G. (1988) in *Plant Transposable Elements*, ed. Nelson, O. (Plenum, New York), pp. 137-148.

21. Brown, J. J., Mattes, M. G., O'Reilley, C. & Shepherd, N. S. (1989) *Mol. Gen. Genet.* **215,** 239-244.

22. Tsai, C. Y. & Nelson, O. E. (1966) *Science* **151,** 341-343.

23. Xu, X. J., Hsia A.-P., Zhang L., Nikolau, B. J. & Schnable, P. S. (1995) *The Plant Cell* **7,** 2151-2161.

24. Schwarz-Sommer, Z., Shepherd, N., Tacke, E., Gierl, A., Rhode, W., Leclercq, L., Mattes, M., Berndtgen, R., Peterson, P. & Saedler, H. (1987) *EMBO J.* **6,** 287-294.

25. Bhave, M. R., Lawrence, S., Barton C. & Hannah, L. C. (1990) *Plant Cell* **2,** 581-588.

26. Chen, M. & Bennetzen, J. L. (1996) *Plant Mol. Biol.* **32,** 999-1001.

27. Chen, M., SanMiguel, P., Bennetzen, J. L. (1998) *Genetics* **148,** 435-443.

28. Wessler, S. R., Bureau, T. E. & White, S. E. (1995) *Curr. Opin. Genet. Dev.* **5**, 814-821.

29. Brown, J. & Sundaresan, V. (1991) *Theor. Appl. Genet.* **81,** 185-188.

30. Timmermans, M. C. P., Das, O. P. & Messing, J. (1996) *Genetics* **143,** 1771-1783.

31. Ohta, K., Shibata, T. & Nicolas, A. (1994) *EMBO J.* **13,** 5754-5763.

32. Wu, T. C. & Lichten M. (1994) *Science* **263,** 515-518.

33. Fu, H., Park, W., Yan, X., Zheng, Z., Shen, B. & Dooner, H. K. (2001) *Proc. Natl. Acad. Sci. USA* **98,** 8903-8908.

34. Dooner, H. K. & Martinez-Ferez, I.M. (1997) *Plant Cell* **9,** 1633-1646.

35. Okagaki, R. J. & Weil, C. F. (1997) *Genetics* **147,** 815-821.

36. Patterson, G. I., Kubo, K. M., Shroyer T. & Chandler, V. L. (1995) *Genetics* **140,** 1389-1406.

37. Eggleston, W. B., Alleman, M. & Kermicle, J. L. (1995) *Genetics* **141,** 347-360.

38. Borts, R. H. & Haber, J. E. (1989) *Genetics* **123,** 69-80.

39. Colot, V., Maloisel, L. & Rossignol, J.-L. (1996) *Cell* **86,** 855-864.

40. Dooner, H. K. (1986) *Genetics* **113,** 1021-1036.

41. Hilliker, A. J., Harauz, G., Reaume, A. G., Gray, M., Clark S. H. & Chovnick, A. (1994) *Genetics* **137,** 1019-1026.

42. Yeadon, P. J. & Catcheside, D. E. A. (1998) *Genetics* **148,** 113-122.

Table 1. No. of recombinants isolated

	1992		1993	
	Clsh*	clrd*	Clsh	clrd
No. isolated	30	108	48	57
No. tested	19	64	22	45
No. confirmed†	18	28	21	43
Corrected No.‡	28	47	46	54

*Clsh: colored shrunken kernels; clrd: colorless round kernels.

†Nine of the confirmed recombinants were not analyzed because homozygotes were not available.

‡Corrected no. = No. isolated x (No. confirmed/No. tested). Corrected numbers were used to calculate the genetic distance between *a1* and *sh2*.

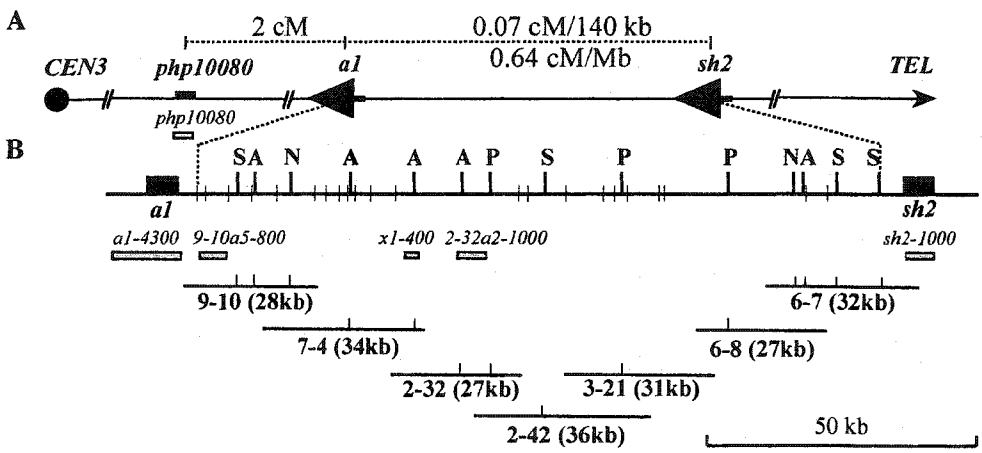


Fig. 1. Yao *et al.*

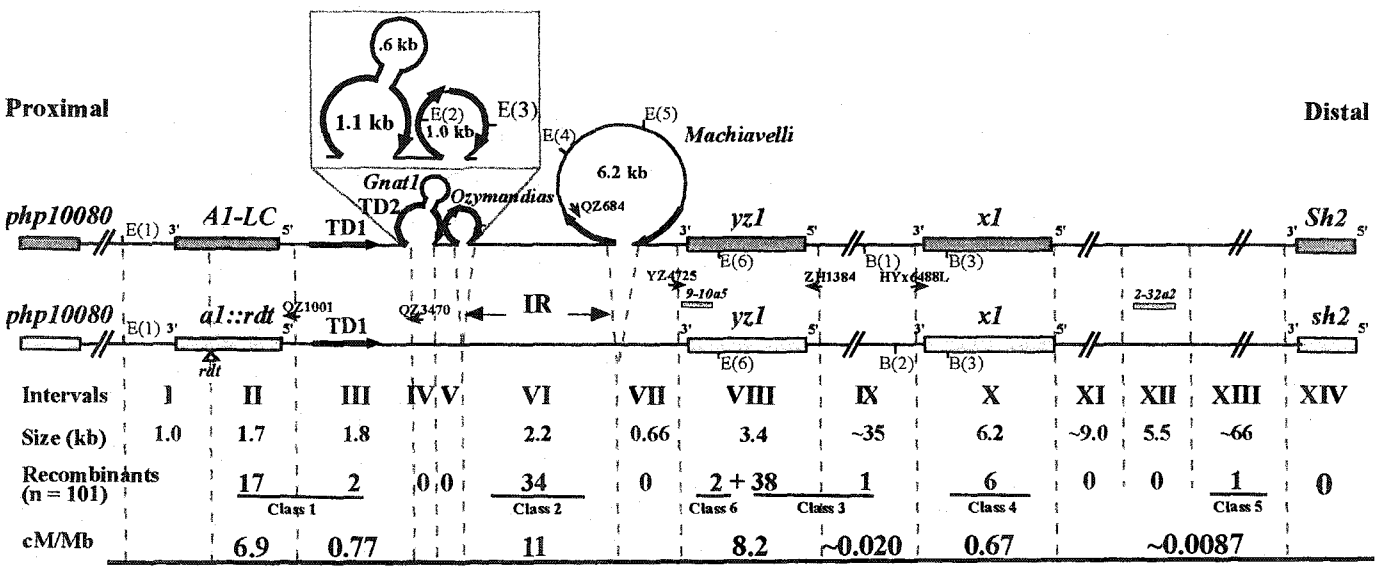


Fig. 2. Yao *et al.*

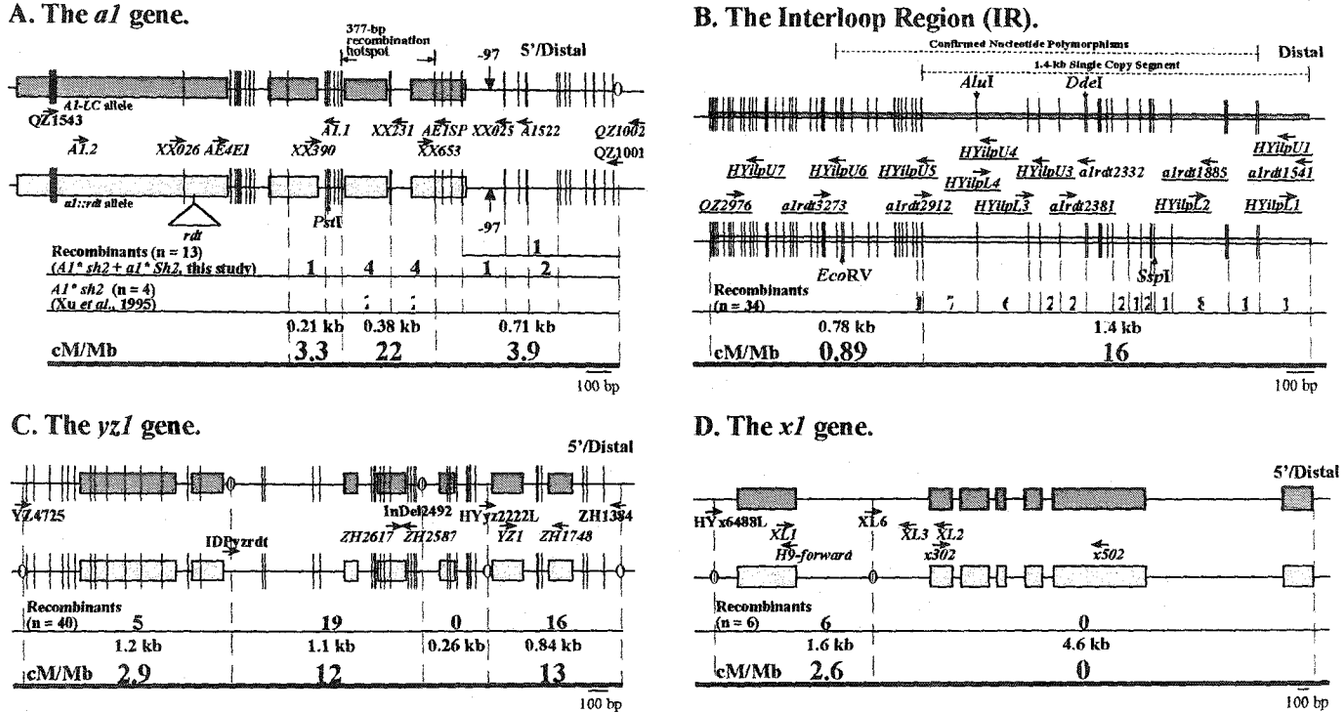


Fig. 3. Yao *et al.*

Supporting Information Published On-Line

The Structures of the *A1-LC Sh2*, *a1-mum2 Sh2*, and *A1-LH82 Sh2* Haplotypes Are Identical or Nearly Identical in the Region between the *a1* and *yz1* Loci. For example, the sequences of the *A1-LC*, *a1-mum2* and *A1-LH82* alleles are identical (GenBank accession nos. X05068, AF363390, AF363391, AF347696, and U46063). In addition, sequence analysis of the *a1-mum2 Sh2* and *A1-LH82 Sh2* haplotypes and PCR analysis of the *A1-LC Sh2* haplotype has established that all three haplotypes contain the 1.1-kb tandem duplication (TD2) that includes the *Gnat1* insertion (Fig. 2). Further, sequence analysis of the *A1-LH82 Sh2* haplotype and extensive DNA gel blotting, PCR and sequencing analyses of the *a1-mum2 Sh2* and *A1-LC Sh2* haplotypes has established that all three contain the *Ozymandias* and *Machiavelli* retrotransposons (unpublished work). There is, however, a single nucleotide polymorphism (SNP) in *Ozymandias* between the *a1-mum2 Sh2* and *A1-LH82 Sh2* haplotypes. Finally, DNA sequencing of portions of the *A1-LC Sh2* haplotype and many of its derivative recombinant haplotypes has established that the *A1-LC Sh2* haplotype is identical to GenBank accession nos. AF347696 and AF434192 at every SNP or small indel that is polymorphic between the *a1::rdt sh2* haplotype and GenBank accession nos. AF347696/AF434192. Hence, a 21,230-bp sequence assembled from GenBank accession nos. AF434192, AF347696, AF363390, X05068, and AF363391 has been designated the "*A1-LC Sh2*" haplotype. Positions 1 - 15,783 of "*A1-LC Sh2*" were derived from positions 1 - 15,783 of GenBank accession no. AF434192; positions 15,784 - 16807 of *A1-LC Sh2* were derived from positions 1,321 - 2,344 of GenBank accession no. AF347696; positions 16,808 - 17,075 of *A1-LC Sh2* were derived from positions 1 - 268 of GenBank accession no. AF363390; positions 17,076 - 20,659 of "*A1-LC Sh2*" were derived from positions 313 -

3,896 of GenBank accession no. X05068; positions 20,660 - 21,209 of "*A1-LC Sh2*" were derived from positions 1 - 550 of GenBank accession no. AF363391; positions 21,210 - 21,230 of *A1-LC Sh2* were derived from positions 4,447 - 4,467 of GenBank accession no. X05068.

Identification of the Maize *x1* Gene. Computational analysis of the sequences of the rice *a1-sh2* intervals revealed a predicted gene (gene *X,* ref. 1). A 2.1-kb rice cDNA clone of gene *X* (ID # R2277) was obtained from the Japanese Rice Genome Research Program (Tsukuba, Ibaraki 305-0854, Japan) and sequenced. Based on the finding that a shotgun plasmid clone (p2-32H9) from the maize *a1-sh2* interval exhibits a high degree of sequence similarity to exon 6 of the rice gene *X*, this rice cDNA (R2277) was used to screen maize cDNA libraries. Three maize cDNAs (2.6, 1.75, and 1.4 kb) were identified that hybridize to the rice gene *X* and serve as templates for PCR using primers designed based on the sequence of clone p2-32H9. The 2.6-kb maize *x1* cDNA clone (X-V1) was isolated from a library prepared from immature tassels of the inbred W22 (2) and was shown to be full length by means of 5' and 3' Rapid Amplification of cDNA Ends (RACE) (GIBCO BRL, Life Technologies) experiments (data not shown); two other clones contain partial *x1* cDNA clones (X-V3 and X-V5) were isolated from a library prepared from seedlings of the inbred CI31A.

The Sequences of the Oligonucleotides Used as Primers for PCR and Sequencing. QZ1543: 5'-AAACATAAAAACAATACGTAATCCAG-3'; A1.2: 5'-GATTGTTGCTT AAGCGCCAATCGT-3'; XX026: 5'-GAGGTCGTCGAGGTGGATGAGCTG-3'; AE4EI: 5'-CGAATTCCGCCAGGGTTTTAGACA-3'; XX390: 5'-TCGGCTTGATTAC CTCATTCT-3'; A1.1: 5'-GTCTTCATTGCACATGCACTGCAC-3'; XX231: 5'-GCC

AAACTCTGATTCGCTCCGTG-3'; XX653: 5'-CGAGCCAGGAGCCGACGAAG-3'; AE1SP: 5'-GACTAGTGCCGGTGCGAGCGAGA-3'; XX025: 5'-GGTAGTTGCAGCG TGTGGTGTT-3'; A1522: 5'-GGGAGTTTGGAGTTGGAGAGG-3'; QZ1001: 5'-GAT ACAGAAGTATATATAAGGGCCAA-3'; QZ1002: 5'-TATTCGTAATGATGTTTAT-3'; QZ3470: 5'-CATCTGAGTGGGAGGCTAAA-3'; QZ2976: 5'-ACTTGTCTCCAT CGCTCT-3'; HYilpU7: 5'-AGACGATTGATGATGATTT-3'; a1rdt3273: 5'-GATTGT CTTTAGGGAACTG-3'; HYilpU6: 5'-GCAGTTCCCTAAAGACA-3'; a1rdt2912: 5'-AACACCCCGCTAACAC-3'; HYilpU5: 5'-GTGTTAGCGGGGTGTT-3'; HYilpL4: 5'-ATCTTGATCCTCTTGAAT-3'; HYilpU4: 5'-CGATGATTCAAGAGG-3'; HYilpL3: 5'-GCTTGCTTGCTTCTGGATGT-3'; HYilpU3: 5'-CAAGCATAAGCATCCATC-3'; a1rdt2381: 5'-TCAACCGTGCTACCAACT-3'; a1rdt2332: 5'-CCGAGTGATAG TAAAGACC-3'; a1rdt1885: 5'-AAAACCAAACGAACATACC-3'; HYilpL2: 5'-ATTCGGTATGTTCGTTTGGTT-3'; HYilpU1: 5'-CAGCCTGTACCAACC-3'; HYilpL1: 5'-CGAAACAGTTACCGAGATAG-3'; a1rdt1541: 5'-CGCTAACTATC TCGGTAACT-3'; QZ684: 5'-GGTTTTTGGGAAGCGTCT-3; YZ4725: 5'-AAATGG TCAGGATAGCTTAGTT-3'; IDPyzrdt: 5'-GAAGTTATGTTCGCGGTG-3'; ZH2617: 5'-CGAACAGGGAAGAATGG-3'; ZH2587: 5'-GCCTGGTTAGCGAAGTTG-3'; HYyz2222L: 5'-CGCCAAAAAAAAAAAAACA-3'; YZ1: 5'-GCGGCGTTGCT GCTGTA-3'; ZH1748: 5'-CACATCCCCGTCTCCT-3'; ZH1384: 5'-GCCATCTCTAC TGTTACCTT-3'; HYx6488L: 5'-ATCTGGGGAAGGGTATCT-3'; XL1: 5'-ATGTTC TTCTTTGAGTG-3'; H9-forward: 5'-ATCGAGGATGATGCAAAG-3'; XL6: 5'-AAA ATCCCCTCGCTGTG-3'; XL3: 5'-ATGAGCGGGAGCCTATG-3'; x302: 5'-CTCTCC CATTCTCTTGATTCCT-3'; XL2: 5'-TGTTCAAAGTGGGAGG-3'; x502: 5'-AGG AATAATAGCGGACCACTTG-3'.

References

1. Chen, M. & Bennetzen, J. L. (1996) *Plant Mol. Biol.* **32,** 999-1001.

2. DeLong, A., Calderon-Urrea, A. & Dellaporta, S. L. (1993) *Cell* **74,** 757-768.

Table 2. Rate of recombination within the *a1-sh2* interval

	No. recombinants*				
Year	Clsh	clrd	Total	Population Size	cM†
1992‡	28	47	75	67,000	0.11 ± 0.01
1993	46	54	100	182,000	0.055 ± 0.005
Total§	74	101	175	249,000	0.070 ± 0.005

Clsh: colored shrunken kernels; clrd: colorless round kernels.

*Based on corrected numbers from Table 1.

†Genetic distances were calculated as follows: cM = (no. of recombinants /population size) X 100.

‡These data have been extended from those of Civardi *et al* (1).

§A homogeneity χ^2 test indicated that there was a significant difference in the rates of recombination between these two years. However, because no differences were found between the distributions of the recombination breakpoints, data were combined across years to calculate recombination rates and genetic distances.

Reference:

1. Civardi, L., Xia, Y. J., Edwards, K. J., Schnable, P. S. & Nikolau, B. J. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 8268-8272.

Table 3. RFLP hybridization patterns associated with intergenic recombination events

		Restriction enzymes and probes used to detect RFLPs						
		*Eco*R I			*Bgl*II		*Eco*R I	No. of
Haplotype	Class*	php10080	a1-4300	9-10a5-800	x1-400	2-32a2-1000‡	sh2-1000	events§
35 *A1* sh2*	1a	LC† (9.5)	N† (8.8)	N (8.8)	a1 (3.0)	a1	a1 (>12)	11
	2a	LC (9.5)	LC (6.8)	N (3.9)	a1 (3.0)	a1	a1 (>12)	8
	3a	LC (9.5)	LC (6.8)	LC (4.3)	a1 (3.0)	a1	a1 (>12)	8
	4a	LC (9.5)	LC (6.8)	LC (4.3)	LC (10)	a1	a1 (>12)	5
	5	LC (9.5)	LC (6.8)	LC (4.3)	LC (10)	LC	a1 (>12)	1
	6	a1† (>12)	LC (6.8)	LC (4.3)	a1 (3.0)	a1	a1 (>12)	2
66 *a1* Sh2*	1b	a1 (>12)	N (7.5)	LC (4.3)	LC (10)	LC	LC (3.5)	8
	2b	a1 (>12)	N (9.9)	LC (4.3)	LC (10)	LC	LC (3.5)	26
	3b	a1 (>12)	a1 (9.5)	a1 (9.5)	LC (10)	LC	LC (3.5)	31
	4b	a1 (>12)	ND†	a1 (9.5)	N (3.5)	LC	LC (3.5)	1

Table 3. (continued)

*Pairs of class 1a/1b, 2a/2b, 3a/3b, and 4a/4b represent hybridization patterns that resulted from reciprocal crossover events.
†LC and a1: hybridization patterns indistinguishable from those of the Line C and *a1::rdt sh2* stocks, respectively. N: novel (nonparental) hybridization pattern. ND: no data. Sizes of fragments (in kb) detected by the indicated RFLP markers are shown in parentheses.
‡Multiple hybridization bands were detected in both the Line C and *a1::rdt sh2* stocks with this probe. A 5.5-kb fragment that is specific to Line C and that cosegregates with the *A1-LC Sh2* haplotype was used for mapping purpose.
§One class 2b, one class 3b and one class 4b event were not subjected to hybridization with probe a1-4300. One class 1b event and one class 2b event were subjected to RFLP analysis with probe x1-400 only. All of these events were analyzed by PCR using allele-specific primers that can detect polymorphisms in the *a1*, *yz1*, and *sh2* loci.

CHAPTER 3. *CIS*-EFFECTS ON MEIOTIC RECOMBINATION ACROSS DISTINCT *a1-sh2* INTERVALS IN A COMMON *ZEA* GENETIC BACKGROUND

A paper to be submitted to *Genetics*

Hong Yao and Patrick S. Schnable

ABSTRACT

Genetic distances across the *a1-sh2* interval varied three fold in three near-isogenic stocks that carry structurally distinct teosinte *A1 Sh2* haplotypes (from *Z. mays* spp. *mexicana* Chalco, *Z. mays* spp. *parviglumis* and *Z. luxurians*) and a common maize *a1::rdt sh2* haplotype. In each haplotype over 85% of recombination events resolved in the proximal 10% of the ~130-kb *a1-sh2* interval. Even so, significant differences were observed in the distributions of recombination breakpoints across subintervals among haplotypes. Each of the three previously detected recombination hot spots was detected in at least one of the three teosinte haplotypes and two of these hot spots were not detected in at least one teosinte haplotype. Moreover, novel hot spots were detected in two teosinte haplotypes. Due to the near-isogenic nature of the three stocks, the observed variation in the distribution of recombination events is the consequence of *cis*-modifications. Although generally negatively correlated with rates of recombination/Mb, frequencies of sequence polymorphisms do not fully account for the nonrandom distribution of recombination breakpoints. This study indicates that estimates of linkage disequilibrium must be interpreted with caution when considering whether a gene has been under selection.

INTRODUCTION

Homologous recombination provides physical connections between pairs of homologous chromosomes during meiosis and thereby helps to prevent non-disjunction. In addition, meiotic recombination generates novel haplotypes upon which natural selection can act. Two types of recombination events result from meiotic recombination: reciprocal crossovers (CO) and unidirectional non-crossovers (NCO). Although evidence from yeast has shown that both events are initiated by double-strand breaks (DSB) (reviewed by PAQUES and HABER 1999), these two types of events probably arise via different pathways (ALLERS and LICHTEN 2001; HUNTER and KLECKNER 2001; CLYNE *et al.* 2003). COs are thought to arise via the DSB repair (DSBR) pathway (SZOSTAK *et al.* 1983; CAO *et al.* 1990; SUN *et al.* 1991) which involves the formation of double Holliday junctions (DHJs) following strand invasion; resolution of these DHJs can result in COs. Although NCOs could also arise via this pathway (following an alternative resolution of DHJs), several pieces of evidence suggest that NCO events may instead arise from the synthesis-dependent strand-annealing (SDSA) pathway that does not involve the formation of DHJs (reviewed by PAQUES and HABER 1999; ALLERS and LICHTEN 2001; HUNTER and KLECKNER 2001).

Meiotic recombination does not occur randomly in a genome or across a chromosome. Eukaryotic genomes contain recombination hot and cold spots where the rates of recombination are much higher and lower than average (reviewed by LICHTEN and GOLDMAN 1995; PUCHTA and HOHN 1996; SCHNABLE *et al.* 1998; PETES 2001). Surprisingly, although the DNA sequences of the human and chimp genomes are highly similar, preliminary data suggest that at least some human hot spots are not conserved in chimps (PENNISI 2004). This is consistent with the finding that within a species, the non-

random distribution of meiotic recombination in a genome can be affected by genetic modifiers that regulate in *cis* and *trans* rates and distributions of recombination events. *Cis*-regulation of recombination has been demonstrated in studies from fungi, mammals and plants. In fungi, hot spots are classified as α, β and γ according to the natures of the sequences that cause the hyper-recombination activity (reviewed by PETES 2001). 'α'-hot spots are caused by sequences that are transcription factor binding sites and that require the binding of transcription factors to activate the hot spot. 'β'-hot spots are caused by sequences that are thought to cause the exclusion of nucleosomes resulting in higher accessibility of a region to the recombination machinery. 'γ'-hot spots are associated with sequences with high G+C content. In addition to the natures of sequences within or in the vicinity of a hot spot that can regulate recombination in *cis*, sequence polymorphisms between DNA segments residing on a pair of homologues can affect both recombination rates/Mb and the distribution of recombination events. Both large insertion/deletion polymorphisms (InDels) and high density of small sequence polymorphisms, including single nucleotide polymorphisms (SNPs) and small InDels, reduce recombination rates/Mb in fungi, mammals, and plants (reviewed by MODRICH 1996; SCHNABLE *et al.* 1998; BORTS *et al.* 2000). In *S. cerevisiae*, two small sequence polymorphisms are sufficient to significantly decrease rates of meiotic recombination (BORTS *et al.* 1990).

In maize, characterized *cis*-modifiers of meiotic recombination include heterochromatic centromeres that reduce frequency of crossovers in nearby regions; heterozygous knobs that are heterochromatic have similar effects (CARLSON 1977; RHOADES 1978). Polymorphisms due to chromosome rearrangements caused by large deletions, inversions and translocations also reduce recombination rates/Mb (ROBERTSON 1967, 1984;

PHILLIPS 1969; CARLSON 1977). TIMMERMANS *et al.* (1997) identified a *cis*-factor in the *Sh1-Bz1* interval from the inbred line A188 that increases recombination rates/Mb locally but the nature of this factor has not been defined. Higher-resolution analyses of *cis*-modifiers of meiotic recombination have been performed in genic recombination hot spots of maize. As is true in other species, sequence polymorphisms in maize genes can influence recombination in *cis,* although the impact seems to be significantly less than that in other species. Recombination rates/Mb in the *a1* (XU *et al.* 1995) and *bz1* (DOONER and MARTINEZ-FEREZ 1997) loci are suppressed by non-autonomous transposon insertions. Sequence polymorphisms at the *bz1* locus also affect recombination resolution sites and the ratio of NCO/CO events (DOONER and MARTINEZ-FEREZ 1997; DOONER 2002). The insertion of a *Mu1* transposon at the 5' end of the *a1* gene does not, however, change the pattern of recombination resolution (XU *et al.* 1995). These studies of intragenic recombination have revealed *cis*-modifiers that influence meiotic recombination in maize genes. Nevertheless, absent an analysis of the *cis*-effects on the rates and distribution of recombination across a multigenic interval it is not possible to answer questions such as why genes are more likely to be recombination hot spots than intergenic regions and whether intragenic and intergenic recombination are similarly regulated by *cis*-modifiers.

To answer these above questions, the *a1-sh2* interval was used as a model. This region was selected because: 1) the multigenic *a1-sh2* interval (YAO *et al.* 2002) allows us to compare *cis*-effects on intragenic as well as intergenic recombination; 2) previous characterization of the distribution of recombination events across the *a1-sh2* interval identified an apparently non-genic hot spot and a genic non-hot spot (YAO *et al.* 2002), the

analysis of which will be informative; 3) this interval is defined by two markers, the *a1* and *sh2* genes, which give kernel phenotypes that facilitate the isolation of meiotic recombinants.

In the current study, approximately 500 recombination events were isolated from near-isogenic plants that carried *A1 Sh2* haplotypes extracted from a maize inbred line and three maize relatives (*Z. mays* ssp. *mexicana* Chalco, *Z. mays* ssp. *parviglumis* and *Z. luxurians*) in combination with a common maize *a1 sh2* haplotype. Phylogenic studies suggest that maize arose from *Z. mays* ssp. *parviglumis* approximately 9,000 years ago (MATSUOKA *et al.* 2002), diverged from ssp. *mexicana* approximately 75,000 years ago, and *Zea mays* diverged from *Zea luxurians* approximately 135,000 years ago (HANSON *et al.* 1996). As predicted by these evolutionary relationships, the *A1 Sh2* haplotypes used in this study are structurally diverse. This allowed us to observe the effects of varying levels of sequence divergence on recombination and to identify putative specific *cis*-modifiers that co-segregate with the *a1-sh2* intervals. Rates of recombination/Mb across the *a1-sh2* interval vary among the *A1 Sh2* haplotypes. Similary, the distributions of recombination breakpoints within the *a1-sh2* interval also differ significantly among haplotypes. Each of three hot spots detected in a prior study was detected in at least one of the three teosinte haplotypes and two of these hot spots were not detected in at least one teosinte haplotype. In addition, novel hot spots were detected in two of the teosinte haplotypes. These variations in recombinational activity can be attributed to the *cis*-effects related to the divergent sequences of the *A1 Sh2* haplotypes.

MATERIALS AND METHODS

Maize genetic stocks: The stocks used to produce progenies carrying recombinant *a1 sh2* haplotypes were derived from genetic crosses between three teosinte lines: *Z. mays* ssp. *mexicana* Chalco (Schnable lab Ac#294; Iltis 28620), *Z. mays* ssp. *parviglumis* (Schnable lab Ac#1322-292; Doebley 1993-1994 292) and *Z. luxurians* (Schnable lab Ac#291; Beadle VII.A.4) as well as the maize inbred "Line C" (a color-converted version of W22) and the near-inbred maize *a1::rdt sh2* stock. Like the *A1-LC* allele from Line C, the *A1* alleles derived from teosinte condition colored kernel phenotypes and in this report are designated *A1-mex*, *A1-par* and *A1-lux*. The *a1::rdt* allele conditions a recessive colorless kernel phenotype because the function of the *a1* gene is disrupted by the *rdt* transposon insertion (BROWN *et al.* 1989). The functional *Sh2* alleles derived from teosinte and Line C condition a round kernel phenotype. Kernels homozygous for the mutant *sh2* alleles are shrunken (MAINS 1949; LAUGHNAN 1953; HANNAH and NELSON 1976).

Stocks used to isolate meiotic recombinants were developed by introgressing the *A1 Sh2* haplotypes from the three teosinte lines and maize inbred Line C into the maize *a1::rdt sh2* stock. First, F1 plants were generated from crosses between the maize *a1::rdt sh2* stock and the three teosinte lines as well as Line C. Then a single F1 plant carrying the *A1 Sh2* haplotype from each teosinte and Line C was selected to backcross to the *a1::rdt sh2* stock for 4-5 generations (teosinte) or 10 generations (Line C). In each generation colored round kernels carrying the *A1 Sh2* haplotypes were selected for the next generation of backcrosses. The resulting stocks carry distinct *A1 Sh2* haplotypes in a common genetic background that is derived from the near inbred *a1::rdt sh2* stock and have the genotype of *A1 Sh2/a1::rdt sh2*.

In this manuscript these heterozygous stocks are also referred as the mex, par, lux and LC2 stocks and the corresponding *A1 Sh2* haplotypes as mex, par, lux and LC haplotypes.

Isolation and confirmation of meiotic recombinants and calculation of genetic distance: The mex, par, lux and LC2 stocks were used as female parents (listed first) and the near-inbred *a1::rdt sh2* stock as male parent in genetic crosses, *A1 Sh2/a1::rdt sh2* x *a1::rdt sh2/a1::rdt sh2*, to generate meiotic recombinants (Table 2) following procedures similar to those described previously (CIVARDI *et al.* 1994; XU *et al.* 1995). Kernels from these crosses that exhibit non-parental phenotypes (colored shrunken and colorless round *vs.* parental colored round and colorless shrunken) presumably carry recombinant chromosomes (designated *A1* sh2* and *a1* Sh2*) resulting from meiotic recombination that could be COs between the *a1* and *sh2* loci or NCOs (*e.g.*, gene conversions) at the *a1* or *sh2* loci.

Samples of the putative recombinants from each source were tested via genetic crosses and molecular marker analysis as described previously (XU *et al.* 1995; YAO *et al.* 2002). Based on the frequency of putative recombinants confirmed within each sample, the number of the true recombinants isolated from each cross could be estimated and used to calculate the genetic distance between the *a1* and *sh2* loci in the corresponding female parent. Stocks homozygous for the recombinant haplotypes (*A1* sh2* and *a1** Sh2) were generated as described previously (CIVARDI *et al.* 1994; XU *et al.* 1995) and used to map the recombination breakpoints.

Breakpoints associated with the confirmed recombinants from the LC2 stock were not physically mapped because a detailed analysis of the distribution of recombination breakpoints associated with the LC haplotype has been conducted previously using a

different stock (referred as the LC1 stock in this manuscript) that carries the same *A1 Sh2* and *a1::rdt sh2* haplotypes as the LC2 stock (YAO *et al.* 2002).

Because no significant differences (p-values > 0.05) were observed between the distributions of breakpoints associated with the two classes of recombinants (*A1* sh2 vs. a1* Sh2*), these two classes of recombinants were combined for subsequent analyses.

Sequences of the *A1 Sh2* haplotypes from the three teosinte lines: Portions of the *A1 Sh2* haplotype from Line C (the "*A1-LC Sh2*" haplotype, YAO *et al.* 2002) (GenBank accessions AF434192, AF347696, AF363390, X05068, and AF363391) and the *a1::rdt sh2* haplotype (GenBank accession AF072704) have been sequenced previously. To sequence the corresponding regions of the three teosinte *A1 Sh2* haplotypes used in this study, plants with the genotype *A1 Sh2* (teosinte)/*a1::rdt sh2* were self pollinated. Colored and round kernels were planted. DNA samples isolated from plants that are homozygous for the teosinte *A1 Sh2* haplotypes were PCR amplified using primers from the *a1*, *yz1* loci and the the Interloop Region between the two loci (Figure 2A). Purified PCR products were then sequenced directly.

The 11-kb *a1-yz1* interval from *Z. mays* ssp. *mexicana* Chalco (GenBank accession AY662984) was assembled from sequences of eight overlapping PCR fragments that ranged in size from about 1 to 3.5 kb. Results obtained from RFLP analyses using probes derived from the *a1* and *yz1* loci and partial sequencing of the amplified product from long range PCR conducted using primers that anneal to the *a1* and *yz1* loci confirmed the organization of the assembled sequence of the 11-kb *a1-yz1* interval (data not shown). The 6.4-kb *a1-yz1* interval from *Z. luxurians* (GenBank accession AY662985) was assembled from sequences of five overlapping PCR fragments that ranged in size from about 0.5 to 2.5 kb, one of which

includes the entire intergenic region and overlaps both the *a1* and *yz1* loci. The entire *a1-yz1* interval from Z. *mays* ssp. *parviglumis* could not be amplified. A 3.9-kb sequence (GenBank accession AY662986) from *yz1* to the distal portion of the Interloop Region was assembled from the sequences of four overlapping PCR fragments of about 0.25 to 1.7 kb. Another 2.3-kb sequence (GenBank accession AY662987) including part of *A1-par* and its 5' upstream region was assembled from two overlapping PCR fragments of about 1.1 and 1.5 kb. The region between these two sequenced segments could not be PCR amplified.

Portions (part of the exon 2 to part of exon 7) of the three teosinte *X1* alleles were also PCR-amplified and sequenced. For each of the three *X1* alleles (GenBank accessions AY656756-AY656758), sequences (3.6 kb for *X1-mex* and *X1-par* and 3.3 kb for *X1-lux*) were assembled from three overlapping PCR fragments of about 1.5 (for *X1-mex* and *X1-par*) or 1.3 (for *X1-lux*) to 1.8 kb.

Oligonucleotides for PCR and sequencing: Sequence comparisons between the three teosinte *A1 Sh2* haplotypes and the *a1::rdt sh2* haplotype revealed many polymorphisms, including SNPs and InDels, which can be used as markers to map the recombination breakpoints. Oligonucleotides were designed based on sequences from the three teosinte *A1 Sh2* haplotypes as well as the maize *a1::rdt sh2* haplotype. Details regarding these primers, including their haplotype specificities, are presented in Table 1. These primers were used for PCR amplification and sequencing to map the recombination breakpoints relative to sequence polymorphisms that exist between the maize *a1::rdt sh2* haplotype and the three teosinte *A1 Sh2* haplotypes.

Statistical methods: Homogeneity χ^2 tests were used to compare genetic distances/recombination rates per Mb between the *a1* and *sh2* loci among the mex, par, lux

and LC haplotypes (Table 2, Figure 1A). In these tests, the corrected numbers of recombinants and population sizes from each stock were used (Table 2). The rates of recombination/Mb in each of the subintervals defined by sequence polymorphisms (Figures 2D, 3E, 5D) were also compared among different teosinte *A1 Sh2* haplotypes. Because not all the recombinants between the *a1* and *sh2* loci could be mapped (*e.g.*, some were not recovered), the sizes of populations that correspond to the numbers of mapped recombinants were calculated using the formula: Actual population size x (Number of mapped recombinants / Number of corrected recombinants). The numbers of mapped recombinants and their corresponding population sizes were then used in the homogeneity χ^2 test. These calculated population sizes were also used to obtain expected numbers of recombinants in each subinterval assuming that the rate of recombination/Mb across the *a1-sh2* interval was equal to the genome's average (2.1 cM/Mb). Then the expected and actual numbers of recombinants mapped to a subinterval as well as the corresponding calculated population size were used in the goodness-of-fit χ^2 test to compare the observed rate of recombination/Mb in a subinterval to the genome's average. Via a similar approach, the observed rate of recombination/Mb in a subinterval was compared to the average rate of recombination/Mb between the *a1* and *sh2* loci using the goodness-of-fit χ^2 test. The distributions of recombination breakpoints in a given subinterval from different teosinte haplotypes were compared via the χ^2 contingency test. These distributions were also compared to the expected patterns obtained under the null hypothesis that recombination events resolve randomly in a given subinterval via the χ^2 contingency test. The Freeman-Halton test (FREEMAN and HALTON 1951) was used to check the reliability of the χ^2 and p values for subintervals that contain fewer than five recombination breakpoints. The Freeman-Halton

test conducts multiple permutations of the original data to estimate the chance of obtaining a χ^2 value that is equal to or greater than the value from the original χ^2 contingency test. χ^2 values and the resulting p values obtained from the original tests were considered reliable if the chance calculated by the Freeman-Halton test (10,000 permutations) is less than 0.05. All χ^2 contingency tests reported as being statistically significant had Freeman-Halton p values of less than 0.05.

The "frequency of sequence polymorphisms" was calculated as the absolute number of polymorphisms (counting each SNP and InDel one time) between a given *A1 Sh2* haplotype and the *a1::rdt sh2* haplotype per 100 bp of the *a1::rdt sh2* haplotype. The correlation coefficient of the frequencies of sequence polymorphisms and rates of recombination/Mb were calculated across all three haplotypes. For these calculations, data from subintervals I-1, I-2, II, III, IV-1, IV-2, IV-3 and VI (Figures 2D-E, 3E-F and 5D-E) in that haplotype were pooled. The significance of the correlation coefficient was determined using Student's *t*-tests. A conservative estimate of the frequency of sequence polymorphisms in the partially sequenced subinterval III-par was obtained by dividing the number of sequence polymorphisms in the sequenced portion by the entire length of this subinterval in the *a1::rdt sh2* haplotype.

RESULTS

Recombination rates/Mb between the *a1* and *sh2* loci differ among haplotypes: To characterize *cis*-effects on meiotic recombination across the *a1-sh2* interval, near-isogenic mex, par, lux and LC2 stocks that carry distinct *A1 Sh2* haplotypes (referred as mex, par, lux and LC haplotypes, respectively) from three teosinte lines, *Z. mays* ssp. *mexicana* Chalco, *Z.*

mays ssp. *parviglumis* and *Z. luxurians* and the maize inbred Line C were developed (Methods). Meiotic recombinants from each stock were isolated and confirmed (Methods). The genetic distances between the *a1* and *sh2* loci varied approximately 3-fold from 0.065 ± 0.0035 cM in the lux haplotype to 0.20 ± 0.012 cM in the mex haplotype (Table 2). The resulting average rates of recombination/Mb across the *a1-sh2* intervals of these distinct haplotypes range from 0.50 to 1.5 cM/Mb (Figure 2D). Based on a homogeneity χ^2 test the rate of recombination/Mb in the mex haplotype is significantly different from all three others (Figure 1A). The par haplotype exhibits a recombination rate/Mb that is significantly different from that of the lux but not of the LC haplotype. The recombination rates/Mb in the lux and LC haplotypes do not differ significantly.

Structure of the *a1-sh2* interval: The sequences of the three teosinte haplotypes differ from each other and from the LC and *a1::rdt sh2* maize haplotypes by both large InDels and numerous small InDels and SNPs (Figures 2A, 2E). The *a1-sh2* interval was divided to seven subintervals relative to the sequence polymorphisms between the maize *a1::rdt sh2* haplotype and the three teosinte *A1 Sh2* haplotypes (Figure 2A). Subinterval I consists of the 5' two-thirds of the transcribed region of the *a1* gene. Subinterval II contains the *a1* promoter. Subinterval III consists of the intergenic region between the *a1* and *yz1* genes. Subinterval IV contains the entire transcribed region of the *yz1* gene. Subinterval V consists of the intergenic region between the *yz1* and *x1* genes. Subinterval VI contains the 3' end of the transcribed region of the *x1* gene. Subinterval VII contains the 5' end of the *x1* gene and the intergenic region between the *x1* and *sh2* genes. For each teosinte haplotype, the frequencies of sequence polymorphisms (Methods) between the *A1 Sh2* and *a1::rdt sh2*

haplotypes vary across the subintervals (Figure 2E). Within the same subinterval, frequencies of sequence polymorphisms also differ among haplotypes.

Mapping breakpoints associated with meiotic recombinants across the *a1-sh2* interval: The recombination breakpoints associated with 99% of the confirmed recombinants (Table 2) from the mex (176/177), par (106/106) and lux (183/185) stocks were mapped to the seven subintervals relative to these sequence polymorphisms (Figure 2B). For each recombinant haplotype, only one breakpoint was detected between the *a1* and *sh2* loci suggesting that most recombinant haplotypes resulted from simple recombination events (i.e., without mosaicism).

The distributions of recombination breakpoints across the *a1-sh2* interval differ among haplotypes: In each of the three distinct teosinte *A1 Sh2* haplotypes the distribution of recombination breakpoints is significantly different than expected based on the null hypothesis of a random distribution across the *a1-sh2* interval (p-values $< 2.2e^{-16}$) (Figure 2B *vs.* 2C). In each of the three haplotypes, over 85% of the breakpoints mapped to the *a1-yz1* region (subintervals I-IV, Figure 2B), even though this region comprises less than 10% of the length of the entire *a1-sh2* interval (Figure 2A). Consistent with prior studies conducted using the maize LC and *a1::rdt sh2* haplotypes (YAO *et al.* 2002), most of the recombinants that map to the remainder of the *a1-sh2* interval (i.e., subintervals V-VII) from each of the three teosinte haplotypes, map to the 3' end of the *x1* gene (i.e., subinterval VI) or 5' of the coding region of *yz1*. Even though general patterns of recombination are conserved across haplotypes, based on χ^2 contingency tests the distributions of recombination breakpoints across the *a1-sh2* interval differ significantly among the three teosinte haplotypes (p-value $< 2.2e^{-16}$). These differences exist between any two of the three haplotypes (Figure 1B), *e.g.*,

within each pair of the teosinte haplotypes the distribution of hot and/or cold spots differs (Table 3, Figures 2B, 2D). As shown in the analyses below, even though all seven subintervals of the three *A1 Sh2* haplotypes have divergent sequences, some subintervals are recombination hot spots in all three haplotypes; some are hot spots in only one or two haplotypes; and some are cold spots in all haplotypes. Hot or cold spots can be defined relative to the *a1-sh2* interval or to the entire genome. In a given stock, regions that exhibit significantly higher or lower recombination rates/Mb than the entire genome's average [2.1 cM/Mb, calculated according to the physical size of ~2,500 Mb (ARUMUGANATHAN and EARLE 1991) and genetic size of 5,289 cM (DAVIS, personal communication, cited in YAO *et al.*, 2002) for the maize genome] are defined as global hot or cold spots; regions that exhibit recombination rates/Mb that are significantly higher or lower than the *a1-sh2* interval within the corresponding haplotype are defined as local hot or cold spots; regions that are none of above are considered average spots (Table 3).

Not all genes are hot spots and *cis*-modifiers can convert a genic hot spot to an average spot: The transcribed regions of most maize genes that have been characterized are recombination hot spots (reviewed by SCHNABLE *et al* 1998). Even so, YAO *et al.* (2002) found that the transcribed region of the *x1* gene in the *a1-sh2* interval associated with the LC haplotype is not a recombination hot spot, thereby establishing that not all genic regions are hot spots in the maize genome.

To test whether *cis*-modifiers affect the recombination activity of genic regions in the *a1-sh2* interval, the rates of recombination/Mb within each genic region in each haplotype were examined. The *x1* gene is located in subintervals VI and VII (Figure 2A). Subinterval VI consists of the 3' portion of the *x1* locus. Rates of recombination/Mb based on the

observed recombination breakpoints mapped to subintervals VI-mex and VI-lux are significantly higher than the average rates of recombination/Mb across the corresponding *A1 Sh2* haplotypes (Figure 2B-D, Table 3). The observed rate of recombination/Mb in subinterval VI-par, however, is not significantly different than the average rate of recombination/Mb in the par haplotype (Figure 2B-D, Table 3). The rates of recombination/Mb in subinterval VI in the three haplotypes are not significantly different than the genome's average (Figure 2D, Table 3). Therefore, subinterval VI-mex and VI-par are local recombination hot spots; subinterval VI-par is an average spot.

The 5' portion of the *x1* gene is located in subinterval VII (Figure 2A). Even if all the recombination breakpoints that occurred within the ~85 kb subinterval VII (Figure 2B) map to within the transcribed region of the *x1* locus located in subinterval VII, the 5' transcribed region of *x1* would be an average spot in the mex haplotype and global cold spots in both the par and lux haplotypes (data not shown). Correspondingly, rate of recombination/Mb in the entire transcribed region of the *x1* locus is only 2.9 cM/Mb in the mex haplotype, 0.47 cM/Mb in the par haplotype, and 0.96 cM/Mb in the lux haplotype. These rates are equivalent to (p-value = 0.52) or significantly less than (p-values < 0.030) the genome's average (2.1 cM/Mb) and are not significantly different than expected if the distributions of breakpoints were random across the *a1-sh2* intervals of all three haplotypes (p-values > 0.16). Therefore, consistent with previous studies using the maize LC haplotype (YAO *et al.* 2002), in none of the teosinte haplotypes is the *x1* gene as a whole a recombination hot spot. Indeed, in the par and lux haplotypes the *x1* gene is a global cold spot.

The transcribed region of the *yz1* gene is a local and global hot spot in the LC haplotype (YAO *et al.* 2002). As discussed earlier, the majority of recombinants from the

mex, par and lux stocks (49%, 63% and 72%, respectively) resolved in the transcribed region of *yz1*, subinterval IV (Figure 2B). This resulted in rates of recombination/Mb in subinterval IV that differ significantly from the average rates of recombination/Mb across the corresponding *A1 Sh2* haplotypes (Figure 2B-D, Table 3). The rates of recombination/Mb are 31 cM/Mb, 20 cM/Mb and 15 cM/Mb in subintervals IV-mex, IV-par and IV-lux, respectively, values that are significantly higher than the genome's average and significantly different from each other (Figure 2D, Table 3). These results establish that the transcribed region of the *yz1* locus (subinterval IV) is a local and global recombination hot spot in each of the distinct teosinte haplotypes.

The transcribed region of the *a1* gene is also a recombination hot spot in the LC haplotype (CIVARDI *et al.* 1994; XU *et al.* 1995; YAO *et al.* 2002). This region corresponds to subinterval I in the current study (Figure 2A). Breakpoints associated with 19% and 11% of the recombinants obtained from the par and lux stocks map to subinterval I, which resulted in significantly higher rates of recombination/Mb than the average rates of recombination/Mb in each *A1 Sh2* haplotype (Figure 2B-D, Table 3). The recombination rates/Mb in subinterval I-par and subinterval I-lux are 19 cM/Mb and 7.1 cM/Mb, values that are significantly higher than the genome's average (Figure 2D, Table 3). Hence, subinterval I-par and subinterval I-lux are both local and global recombination hot spots. In contrast, only 2.8% of the recombinants from the mex stock resolved in subinterval I. The rate of recombination/Mb in this subinterval is not significantly different from the average rate across the mex haplotype and the genome's average (Figure 2B-D, Table 3). Hence, subinterval I-mex is an average spot (Figure 2D, Table 3).

Based on the existence of transcription factor binding sites between positions –130 and +1, subinterval II (GROTEWOLD *et al.* 1994; TUERCK and FROMM 1994) contains the *a1* promoter. Breakpoints associated with 21% of the recombinants from the mex stock mapped to subinterval II-mex. The corresponding rate of recombination/Mb subinterval II-mex is significantly higher than the average recombination rate/Mb of the mex haplotype (Figure 2B-D, Table 3). In addition, subinterval II-mex exhibited a recombination rate/Mb (59 cM/Mb) that is significantly higher (about 30-fold) than the genome's average (Figure 2D, Table 3). Therefore, subinterval II-mex is both a local and global intergenic recombination hot spot. Significantly, subinterval II-mex has no overlap with the 377-bp genic *a1*-hot spot identified in the maize LC haplotype (Figure 3A-B; XU *et al.* 1995; YAO *et al.* 2002). In contrast to what is observed in subinterval II-mex, breakpoints associated with only 4.7% and 2.2% of the recombinants from the par and lux stocks, respectively, mapped to subintervals II, respectively. The resulting rates of recombination/Mb are not significantly different from the average rates of recombination/Mb across these two *A1 Sh2* haplotypes and the genome's average (Figure 2B-D, Table 3).

These analyses of the *a1* gene suggest that *cis*-modifiers associated with the sequence divergence among the three *A1 Sh2* haplotypes can convert both a transcribed genic hot spot (i.e., subinterval I in the par and lux haplotype) and an un-transcribed genic hot spot (*e.g.*, subinterval II-mex) into average spots (i.e., subinterval I-mex and subintervals II-par and II-lux).

Not all intergenic regions are cold spots and *cis*-modifiers can convert a non-genic cold spot into a hot spot: It has been hypothesized that almost all meiotic

recombination events in eukaryotic genomes occur in genes (THURIAUX 1977). This hypothesis therefore predicts that intergenic regions are recombination cold spots.

Characterization of the maize *a1-sh2* interval did not find evidence for the presence of genes other than *a1*, *yz1*, *x1* and *sh2* (YAO *et al.* 2002). Similar analyses of the rice and sorghum *a1-sh2* intervals also failed to identify other genes (CHEN and BENNETZEN 1996; CHEN *et al.* 1998). Hence, subintervals III, V and most of subinterval VII are thought to be solely intergenic (Figure 2A). Consistent with THURIAUX's hypothesis, in all three teosinte haplotypes subintervals V and VII are local and global recombination cold spots.

In contrast, subinterval III is not a recombination cold spot in any of the three teosinte haplotypes (Figure 2, Table 3). Subinterval III contains a segment (the Interloop Region, Figure 2A) that is a recombination hot spot in the maize LC haplotype (YAO *et al.* 2002). Breakpoints associated with 16% of the recombinants isolated from the mex stock mapped to subinterval III, which resulted in a significantly higher rate of recombination/Mb than the average rate of recombination/Mb across the mex haplotype (Figure 2B-C, Table 3). The recombination rate/Mb in subinterval III-mex is 6.7 cM/Mb, a value that is significantly higher (three fold) than the genome's average and significantly higher than the rates observed in subintervals III-par and III-lux (Figure 2D, Table 3). Breakpoints associated with only 2.8% of the recombinants from the par stock mapped to subinterval III and none of the recombinants from the lux stock resolved in subinterval III (Figure 2B). Whereas the rate of recombination/Mb in subinterval III-par is not significantly different from the average of the par haplotype and the genome's average, the rate of recombination/Mb in subinterval III-lux is significantly lower than both the lux average and the genome's average (Figure 2B-D, Table 3). These results establish that subinterval III is a local and global recombination cold

spot in the lux haplotype, an average spot in the par haplotype, and a local and global hot spot in the mex haplotype. Hence, *cis*-modifiers associated with sequence divergence among the *A1 Sh2* haplotypes are able to convert an intergenic cold spot to a hot spot.

Distributions of recombination breakpoints across the *a1* and *yz1* loci differ among haplotypes: Within maize genes, the distributions of recombination breakpoints differ. In some genes, breakpoints are randomly distributed; in others they are distributed non-randomly (reviewed by SCHNABLE *et al.* 1998). In the *bz1* locus, the presence of SNPs and InDels alters the distribution of recombination breakpoints (DOONER and MARTINEZ-FEREZ 1997). In contrast, although a large InDel caused by a transposon insertion in the *a1* locus (position –97) decreases the rate of recombination/Mb within this gene, it does not affect the distribution of recombination breakpoints (XU *et al.* 1995). To better understand the effects of sequence polymorphisms on patterns of intragenic recombination, the distributions of recombination breakpoints that resolved within the *a1* (subintervals I-II) and *yz1* (subintervals VI-V-1) genes from each of the three near-isogenic stocks were compared to each other and to data from the LC haplotype previously characterized by YAO *et al.* 2002 (Figures 3-5).

The a1 locus: Using the IDP primer, aIDPrdt4, recombinants from the mex, par and lux stocks with breakpoints in subinterval I could be mapped to two smaller subintervals (I-1 and I-2, Figure 3). Subinterval I-2 contains the 377-bp recombination hot spot previously identified in the LC haplotype (XU *et al.* 1995; YAO *et al.* 2002). The distribution of recombination breakpoints derived from the lux haplotype does not differ significantly from that expected if recombination occurs randomly across the *a1* locus (Table 4). In contrast, the distributions associated with the other three haplotypes do differ significantly from

random (Table 4). In the par and LC haplotypes, recombination breakpoints clustered in subinterval I-2; in the mex haplotype they clustered in subinterval II. Significant differences were observed in the distributions of recombination breakpoints among most of the haplotypes (Figure 1C).

The yz1 locus: Recombination breakpoints derived from the mex, par and lux stocks that resolved in subintervals IV and V were mapped to higher resolution using the haplotype-specific primers indicated in Figure 4. Subintervals IV-1, IV-2 and IV-3 contain the entire coding region of the *yz1* gene and subinterval V-1 contains the ~200-400 bp upstream of the beginning of the *yz1* coding region. The LC haplotype was not included in this analysis because the *yz1* markers that are polymorphic between the *a1::rdt sh2* haplotype and all of the teosinte *A1 Sh2* haplotypes are monomorphic between the *a1::rdt sh2* and the LC haplotypes. Across all of subinterval IV, no significant differences were observed in the distributions of recombination breakpoints between the par and mex haplotypes, but the distributions in both of these haplotypes differ significantly from that of the lux haplotype (Figure 1D). This is caused by the significantly lower rate of recombination/Mb in subinterval IV-1-lux as compared to the corresponding intervals of the par and mex haplotypes (Figure 4D).

These high-resolution mapping experiments demonstrated that *cis*-modifiers can alter the patterns of distribution across both of the analyzed two genes.

Distributions of recombination breakpoints across an intergenic region differ among haplotypes: Subinterval III consists of the intergenic region between the *a1* and *yz1* genes. Prior analyses of this region revealed that the LC haplotype contains two large retrotransposon insertions that are not present in the *a1::rdt sh2* haplotype (YAO *et al.* 2002).

The 2.2-kb between these two insertions is termed the "Interloop Region" (IR) in the LC haplotype. The 800-bp proximal portion of the IR consists of repetitive sequences. The 1.4-kb distal portion of the IR (Figure 5) is an apparently non-genic, single-copy recombination hot spot.

Subinterval III is structurally very polymorphic among haplotypes (Figure 2 A). Much of the IR has been deleted from subinterval III-lux. Even though ~900 bp of the 1.4-kb single-copy distal portion of the IR has been retained, no recombinants occurred in any portion of subinterval III-lux. It was not possible to sequence all of subinterval III-par, but this haplotype retains at least 900-bp of the 1.4-kb single-copy distal portion of the IR. Even so, this region is not a recombination hot spot in the par haplotype.

In contrast, subinterval III-mex, which is structurally similar to that of the *a1::rdt sh2* haplotype, is both a local and global recombination hot spot. Recombination breakpoints from subinterval III-mex were mapped to higher-resolution via PCR and sequencing (Figure 5). In contrast to what is observed in subinterval III-LC (YAO *et al.* 2002), the distribution of recombination breakpoints across subinterval III-mex is not significantly different from a random pattern (p-value = 0.27).

DISCUSSION

The highly polymorphic intergenic region between the *a1* and *yz1* loci among teosinte and maize haplotypes: Sequence comparisons of large multigenic intervals among maize haplotypes revealed noncollinearities in both genic (FU and DOONER 2002; SONG and MESSING 2003) and non-genic (FU and DOONER 2002; YAO *et al.* 2002; SONG and MESSING 2003) regions. This study extends these sequence comparisons of multigenic haplotypes to

teosinte. The intergenic region (subinterval III, Figure 2A) between the *a1* and *yz1* genes is highly polymorphic among the maize and teosinte haplotypes. Subinterval III ranges in size from ~1.1 kb in the teosinte lux haplotype to ~13 kb in the maize LC haplotype (YAO *et al.* 2002). This intergenic region is ~5 kb in the maize *a1::rdt sh2* and teosinte mex haplotypes. The expansion of this region in the LC haplotype is caused by transposon and retrotransposon insertions. The reduction of this interval may be caused by deletion events. Although maize arose from *Z. mays* ssp. *parviglumis* (MATSUOKA *et al.* 2002), over the entire *a1-yz1* region the mex haplotype is more similar to the maize *a1::rdt sh2* haplotype than the par haplotype (Figure 2E). This is consistent with the view that gene flow from ssp. *mexicana* has contributed substantially to the maize gene pool after domestication (MATSUOKA *et al.* 2002).

Sequence polymorphisms have *cis*-effects on meiotic recombination across the *a1-sh2* interval: The amount, type and distribution of sequence polymophisms between each of the four maize and teosinte *A1 Sh2* haplotypes (LC, mex, par and lux) and the *a1::rdt sh2* haplotype differ dramatically (Figure 2A, E). There are also significant differences in the rates of recombination/Mb that occur across the *a1 -sh2* interval in these four *A1 Sh2* haplotypes (Figure 1A, Table 2). In addition, the distributions of recombination breakpoints within the *a1-sh2* interval vary significantly among the three teosinte haplotypes (Table 3, Figure 1B). Because recombination rates/Mb and distribution patterns were characterized in near-isogenic stocks in which each haplotype was paired with a common *a1::rdt sh2* haplotype, we conclude that the sequence polymorphisms that exist among the *A1 Sh2* haplotypes alter recombination in the *a1-sh2* interval.

The overall pattern of recombination across the *a1-sh2* interval is conserved among the diverse haplotypes analyzed in this study. It was found previously that in the LC haplotype the bulk of recombination occurs in the *a1-yz1* interval that comprises ~10% of the physical distance between *a1* and *sh2* loci (YAO *et al.* 2002). Similar patterns were observed in the three teosinte haplotypes analyzed in the current study. Although this large-scale pattern of recombination was conserved across the haplotypes, significant differences were observed in the distributions of recombination breakpoints across subintervals. It was previously established that the *a1-sh2* interval of the LC haplotype contains three recombination hot spots, the transcribed region of *a1*, the 1.4-kb single-copy, proximal region of the IR, and the transcribed region of *yz1* (YAO *et al.* 2002). Although each of the three hot spots detected in the LC haplotype was also detected in at least one of the three teosinte haplotypes, two of these hot spots were not detected in at least one haplotype (Figure 2, Table 3). In addition, new hot spots were detected in some of the teosinte haplotypes.

What causes recombination hot spots? It has been hypothesized that the hot spots detected within maize genes are caused by the suppression of recombination in subgenic regions with higher levels of sequence polymorphisms, creating apparent hot spots in subgenic regions that have few polymorphisms (DOONER and MARTINEZ-FEREZ 1997). This hypothesis was developed based on observations at the *bz1* locus, where recombination breakpoints are distributed randomly across the transcribed portion of the *bz1* locus in plants that are heterozygous for nearly identical alleles (DOONER and MARTINEZ-FEREZ 1997), but distributed in a nonrandom fashion in plants that are heterozygous for *bz1* alleles that exhibit higher frequency of polymorphisms (~1/100 bp). Within many organisms, including bacteria, yeast and mouse, recombination between polymorphic templates (i.e., homeologous

recombination) is suppressed, a process in which involves mismatch repair proteins (reviewed by MODRICH 1996; BORTS *et al.* 2000; EVANS and ALANI 2000). This suppression helps prevent deleterious ectopic recombination between repetitive sequences in a genome (reviewed by MODRICH 1996; BORTS *et al.* 2000; EVANS and ALANI 2000). Hence, the polymorphism hypothesis is attractive because it could help explain how a segmentally duplicated genome such as that of maize (HELENTJARIS *et al.* 1988; GAUT and DOEBLEY 1997) can avoid deleterious ectopic recombination between paralogs.

Within a given haplotype, the rates/Mb and distributions of recombination events across the *a1-sh2* interval are at least partially consistent with this hypothesis. In particular, subintervals that exhibit higher recombination rates/Mb than their flanking subintervals also exhibit lower frequencies of nucleotide similarity than their neighbors (Figure 2D-E). This relationship is less clear when comparing non-adjacent subintervals. How well does the sequence polymorphism hypothesis explain the distribution of recombination breakpoints among the various *a1-sh2* haplotypes? The rate of recombination/Mb is highest in the mex haplotype and lowest in the lux haplotype (Figure 2D). Among the teosinte haplotypes, the sequenced portions of the mex and lux haplotypes are least and most, respectively, polymorphic to the *a1::rdt sh2* haplotype (Figure 2E). Considering all haplotypes together, the correlation coefficient of the frequency of sequence polymorphisms and rate of recombination/Mb is –0.44 (p value < 0.025). Hence, the patterns of sequence polymorphisms do not provide a complete explanation for the non-random distribution of recombination breakpoints across haplotypes.

The *yz1* hot spot (subintervals IV and V-1) that was originally detected in the LC haplotype is conserved in all three teosinte haplotypes. One of the interesting features of this

genic hot spot is that recombination breakpoints cluster at the 5' and 3' ends of the gene in all four haplotypes. Consistent with the polymorphism hypothesis, the central portion of *yz1* that experiences lower recombination rates/Mb is also the most polymorphic portion of this gene in all four haplotypes (YAO *et al.* 2002; Figures 4B and 5). In the mex and par haplotypes, the 5' and 3' ends of *yz1* (subintervals IV-1, 3, V-1) exhibit similarly low levels of polymorphism (Figure 4A-B, 4E) and similar rates of recombination/Mb. In contrast, in the lux haplotype, the 3' portion of *yz1* (subinterval IV-1) is more polymorphic and experiences significantly less recombination than the 5' portion (subintervals IV-3, V-1) (Figure 5).

The *a1* hot spot (subinterval I-2) detected in the LC haplotype is conserved in the par and lux, but not mex, haplotypes (Figure 3). On the other hand, a novel hot spot was detected in subinterval II, the *a1* promoter, of the mex haplotype that was not detected in the LC haplotype or either of the other two teosinte haplotypes.

In the par and lux haplotypes, subinterval I-2 is less polymorphic than subinterval II and subinterval I-2, but not subinterval II, is a local and global hot spot in both haplotypes (Figure 2-3, Table 3). On the other hand, in the mex haplotype subinterval I-2 is more polymorphic than subinterval II and in this haplotype subinterval II, but not subinterval I-2, is a local and global hot spot (Figure 2-3, Table 3).

Hence, analyses of the *yz1* and *a1* genes, and considering only single haplotypes, are generally consistent with the polymorphism hypothesis. Even so, the rates of recombination/Mb observed within subintervals do not exhibit a linear relationship with the rates of polymorphisms within the same subintervals. This is probably because certain types of polymorphisms have greater impacts on recombination than do others, and/or there are

interactions among different subintervals within a haplotype that affect recombination rates/Mb.

The data collected on *bz1* (DOONER and MARTINEZ-FEREZ 1997) and *yz1* focused on the transcribed regions of these genes. The analysis of the *a1* hotspot in the mex haplotype extends the relationship between polymorphisms and recombination to a non-transcribed region (subinterval II).

The analysis of recombination in *a1* across haplotypes provides a less clear picture regarding the relationship between polymorphisms and recombination (Figure 3). Subinterval I-2 from the par haplotype has fewer polymorphisms than the corresponding subinterval of the other haplotypes, and also has the highest rate of recombination/Mb which is significantly higher (four times) than that experienced by subinterval I-2-mex. This occurs even though subinterval I-2-par has only one less SNP than subinterval I-2-mex. Similarly, although I-2-mex has fewer polymorphisms than I-2-lux, I-2-lux and I-2-mex have similar rates of recombination/Mb.

Similarly, a correlation between rates of sequence polymorphisms and recombination rates/Mb across haplotypes is not observed in the *x1* gene. Although *x1* is not a recombination hotspot in the LC haplotype, and the 5' end of x1 is not a recombination hot spot in any of the haplotypes, the 3' end of *x1* (subinterval VI) is a local hot spot in the mex and lux haplotype (Figure 2, Table 3). The polymorphism hypothesis would predict that the 3' ends of the *x1*-mex and *x1*-lux alleles would exhibit a higher frequencies of sequence similarity than the 5' ends of these two alleles and the *x1*-mex and *x1-lux* alleles to exhibit higher degrees of sequence similarity to the *x1* allele from the *a1::rdt sh2* haplotype than does the *x1* allele from the par haplotype. Exactly the opposite is observed. The 5' ends of

x1-mex and *x1*-lux are more similar to the *x1* allele derived from the *a1::rdt sh2* haplotype (with 0.2 and 1.2 sequence polymorphisms per 100 bp, respectively) than are the 3' ends (with 0.7 and 3 sequence polymorphisms per 100 bp, respectively). In addition, the 3' ends of the *x1*-mex and *x1*-par alleles are less similar to the *a1::rdt sh2* haplotype than is the corresponding region of the *x1*-par allele (with 0.6 sequence polymorphisms per 100 bp). These results demonstrate that the polymorphism hypothesis cannot by itself explain the distribution of all genic recombination hot spots.

This hypothesis is further weakened by our analysis of an apparently non-genic region. The apparently non-genic subinterval III can be subdivided into four subintervals (Figure 5). In the mex haplotype, but not in the other haplotypes, subinterval III is both a local and global recombination hot spot (Figure 2, Table 3). Even though rates of polymorphisms vary dramatically among these four subintervals, there is no statistical evidence for a non-random distribution of recombination events in this haplotype. For example, although subintervals III-1 and III-4 exhibit similar rates of recombination/Mb (6.4 and 4.9 cM/Mb), they have quite different frequencies of sequence polymorphisms; the 3-kb subinterval III-1 has only a single SNP, while the 0.7 kb subinterval III-4 contains multiple SNPs and InDels relative to the *a1::rdt sh2* haplotype.

Comparisons between the non-genic hot spot in subinterval III-1-mex and the adjacent genic hot in *a1* strengthens the argument against the polymorphism hypothesis. Although the 0.7-kb genic subinterval II-mex (3 SNPs and 1 small InDel) has a higher rate of polymorphisms than the adjacent 3-kb non-genic subinterval III-1-mex (1 SNP), the former has a nine-fold higher rate of recombination/Mb (59 vs. 6.4 cM/Mb, Figure 5). Even so, within genes there often is a relationship between rates of polymorphism and recombination.

Domestication and recombination. Domestication bottlenecks reduce genetic diversity. Consequently, all other factors being equal, genome-wide rates of recombination/Mb would be expected to increase following domestication because in general the frequency of sequence polymorphisms is negatively correlated with the recombination rate/Mb. Such an increase in recombination rate/Mb could impact various evolutionary processes, *e.g.*, faster fixation of agronomically important alleles during domestication (KIMURA and OHTA 1969; WANG *et al.* 1999).

How do polymorphisms suppress recombination? Any model to explain the mechanism by which polymorphisms suppress recombination needs to take into account the finding that small changes in the frequency of polymorphisms can dramatically alter recombination rates/Mb (*e.g.*, subintervals I-2-par vs. I-2-mex) and that the relationship between the rates of polymorphisms and recombination does not apply in all regions (*e.g.*, *x1* and subinterval III).

It has been proposed that polymorphism-mediated suppression occurs at the level of DSB initiation (DOONER and MARTINEZ-FEREZ 1997). The absence of data regarding the distribution of DSB in plants makes it very difficult to test this hypothesis. Even so, the finding that mutations in yeast genes that encode mismatch repair enzymes inhibit the suppression of homeologous recombination (reviewed by MODRICH 1996; BORTS *et al.* 2000; EVANS and ALANI 2000) provides a significant clue. If the suppression of homologous recombination in polymorphic regions of plant genomes is also dependent upon mismatch repair enzymes, then, because the substrates for mismatch repair are produced after DSB initiation, it is unlikely that polymorphism-mediated suppression of recombination occurs at

the level of DSB initiation, but instead occurs by altering the relative outcomes of DSB repair.

Why do recombination events cluster in genes? Based on the observation that among eukaryotes the physical sizes of genomes vary more than the sizes of genetic maps and that the numbers of genes are fairly constant, THURIAUX (1977) hypothesized that recombination events occur primarily within genes. Consistent with this hypothesis, maize genes are usually recombination hot spots (reviewed by PUCHTA and HOHN 1996; SCHNABLE *et al.* 1998) and many of the hot spots in the *a1-sh2* interval are associated with genes.

As discussed above recombination hot spots often exhibit high levels of sequence similarity. Hence, the high level of sequence conservation in genes probably favors the occurrence of recombination in genes. But it has also been observed that repetitive retrotransposon sequences in intergenic regions exhibit low rates of recombination/Mb (YAO *et al.*, 2002) even when these sequences are homozygous (FU *et al.*, 2002). Hence, a low frequency of sequence polymorphisms can not by itself explain the existence of genic hot spots.

Do region-specific chromatin structures affect meiotic recombination? Even though polymorphisms can suppress recombination in some, but not all, intervals, our results also establish that a high rate of sequence similarity is not sufficient to create a recombination hot spot (e.g., *x1* and subinterval I-mex). We hypothesize that the failure of the *x1* gene to act as a recombination hot spot in most haplotypes even though it exhibits low levels of polymorphism could be explained by the presence of local chromatin structure that does not support high rates of DSB initiation. If this is true, then some features of the mex haplotype must alter chromatin structure in the vicinity of the *x1* gene to allow the 3' end of

this gene to function as both a global and local hot spot. Differences in chromatin structure could also explain the nine-fold lower rate of recombination/Mb within the 3-kb non-genic subinterval III-1-mex and the adjacent subinterval II-mex. This is because even though these intervals have similar rates of polymorphism, the later contains the *a1* promoter making it more accessible to recombination machinery than subinterval III, which contains mostly repetitive sequences. Therefore subinterval II-mex may be similar to 'α'-hot spots of yeast (reviewed by PETES 2001). If this is true, it is clear that region-specific chromatin structure is not sufficient to stimulate recombination because subinterval II is not a hot spot in the more polymorphic lux and par haplotypes. Alternatively, the differences in the rates of recombination/Mb in subintervals I and II in the mex haplotype could be the consequence of competition between these two regions for DSB initiation or resolution sites.

***Cis*-modifiers of recombination can affect linkage disequilibrium (LD).** Whole-genome association mapping based on LD is an efficient tool to identify variant alleles of quantitative trait loci. Recombination shapes the genomic pattern of LD (reviewed by GAUT and LONG 2003). In humans, the pattern of LD is correlated with rates of recombination/Mb; high LD blocks with low rates of recombination/Mb are interspersed with recombination hot spots that exhibit rapid decay of LD (reviewed by GOLDSTEIN 2001; RAFALSKI and MORGANTE 2004). The genome-wide pattern of LD in maize may have a similar structure as a consequence of the non-random distribution of meiotic recombination (reviewed by RAFALSKI and MORGANTE 2004). The patterns of LD vary among populations and changes in rates and distributions of recombination caused by genetic modifiers can contribute to this variation (reviewed by RAFALSKI and MORGANTE 2004). For example, in certain maize and teosinte stocks the recombination cold spot between the *a1* and *yz1* loci would be expected to

exhibit a high degree of LD, whereas in other stocks, LD would be expected to decay rapidly in this interval due to the action of *cis*-modifiers that convert this interval from a recombination cold spot to a recombination hot spot.

High rates of LDs across maize genes are often thought to be associated with strong selection (reviewed by GAUT and LONG 2003; RAFALSKI and MORGANTE 2004). This relationship is complicated by the fact that genes do not exhibit consistent rates of recombination/Mb. For example, in several haplotypes the *x1* gene is a recombination cold spot and would therefore be expected to exhibit a high degree of LD regardless of whether or not it has been under selection. The fact that *cis*-modifiers can affect recombination rates/Mb, and hence LD in a haplotype-specific manner, further complicates the relationship between LD and selection. Hence, it is not possible to conclude that just because a gene exhibits a high degree of LD that it has been under selection. Consequently, additional characterization of genetic modifiers of meiotic recombination is likely to enhance our understanding of the genomic patterns of LD and help us better interpret LD data.

ACKNOWLEDGEMENTS

We thank undergraduate students Kenny Tsang, Luke Brunkhorst and Tim Heisel for technical assistance and Dr. Dan Nettleton (Iowa State University) for suggestions regarding statistical analyses. We thank Drs. Bruce Benz (Facultad de Ciencias, Universidad Nacional Autonoma de Mexico) and John Doebley (University of Wisconsin) for teosinte stocks. This research was supported in part by competitive grants from the United States Department of Agriculture - National Research Initiative Program to P.S.S and Basil J. Nikolau (9701407

and 9901579) and to P.S.S. (0101869 and 0300940) and by Hatch Act and State of Iowa funds.

LITERATURE CITED

ALLERS, T., and M. LICHTEN, 2001 Differential timing and control of noncrossover and crossover recombination during meiosis. Cell **106:** 47-57.

ARUMUGANATHAN, K., and E. D. EARLE, 1991 Nuclear DNA content of some important plant species. Plant Mol. Biol. Rep. **9:** 208-218.

BORTS, R.H., S. R. CHAMBERS and M. F. ABDULLAH, 2000 The many faces of mismatch repair in meiosis. Mutat. Res. **451:** 129-150.

BORTS, R.H., W. Y. LEUNG, W. KRAMER, B. KRAMER, M. WILLIAMSON, S. FOGEL and J. E. HABER, 1990 Mismatch repair-induced meiotic recombination requires the *pms1* gene product. Genetics **124:** 573-584.

CAO, L., E. ALANI and N. KLECKNER, 1990 A pathway for generation and processing of double-strand breaks during meiotic recombination in *S. cerevisiae*. Cell **61:** 1089-101.

CARLSON, W. R., 1977 The cytogenetics of corn, pp. 225-304 in *Corn and Corn Improvement*, edited by G. F. SPRAQUE. Am. SOC. of Agronomy, Madison, WI.

CIVARDI, L., Y. XIA, K. J. EDWARDS, P. S. SCHNABLE and B. J. NIKOLAU, 1994 The relationship between genetic and physical distances in the cloned *a1-sh2* interval of the *Zea mays* L. genome. Proc. Natl. Acad. Sci. USA. **91:** 8268-8272.

CHEN, M., and J. L. BENNETZEN, 1996 Sequence composition and organization in the Sh2/A1-homologous region of rice. Plant Mol. Biol. **32:** 999-1001.

CHEN, M., P. SANMIGUEL and J. L. BENNETZEN, 1998 Sequence organization and conservation in sh2/a1-homologous regions of sorghum and rice. Genetics **148:** 435-443.

CLYNE, R. K., V. L. KATIS, L. JESSOP, K. R. BENJAMIN, I. HERSKOWITZ, M. LICHTEN and K. NASMYTH, 2003 Polo-like kinase Cdc5 promotes chiasmata formation and cosegregation of sister centromeres at meiosis I. Nat. Cell Biol. **5:** 480-485.

DOONER, H. K., 2002 Extensive interallelic polymorphisms drive meiotic recombination into a crossover pathway. Plant Cell **14:** 1173-1183.

DOONER, H. K., and I. M. MARTINEZ-FEREZ, 1997 Recombination occurs uniformly within the bronze gene, a meiotic recombination hotspot in the maize genome. Plant Cell **9:** 1633-1646.

EVANS, E., and E. ALANI, 2000 Roles for mismatch repair factors in regulating genetic recombination. Mol. Cell Biol. **20:** 7839-7844.

FREEMAN, G. H., and J. H. HALTON, 1951 Note on an exact treatment of contingency, goodness of fit and other problems of significance. Biometrika **38:** 141-149.

FU, H., Z. ZHENG and H. K. DOONER, 2002 Recombination rates between adjacent genic and retrotransposon regions in maize vary by 2 orders of magnitude. Proc. Natl. Acad. Sci. USA **99:** 1082-1087.

GAUT, B. S., and J. F. DOEBLEY, 1997 DNA sequence evidence for the segmental allotetraploid origin of maize. Proc. Natl. Acad. Sci. USA **94:** 6809-6814.

GAUT, B. S., and A. D. LONG, 2003 The lowdown on linkage disequilibrium. Plant Cell **15:** 1502-1506.

GOLDSTEIN, D. B., 2001 Islands of linkage disequilibrium. Nat. Genet. **29:** 109-111.

GROTEWOLD, E., B. J. DRUMMOND, B. BOWEN and T. PETERSON, 1994 The myb-homologous P gene controls phlobaphene pigmentation in maize floral organs by directly activating a flavonoid biosynthetic gene subset. Cell **76:** 543-553.

HANNAH, L. C., and O. E. NELSON, 1976 Characterization of ADP-glucose pyrophosphorylase from *shrunken-2* and *brittle-2* mutants of maize. Biochem. Genet. **14:** 547-560.

HANSON, M. A., B. S. GAUT, A. O. STEC, S. I. FUERSTENBERG, M. M. GOODMAN, E. H. COE and J. F. DOEBLEY, 1996 Evolution of anthocyanin biosynthesis in maize kernels: the role of regulatory and enzymatic loci. Genetics **143:** 1395-1407.

HELENTJARIS, T., D. WEBER and S. WRIGHT, 1988 Identification of the genomic locations of duplicate nucleotide sequences in maize by analysis of restriction fragment length polymorphisms. Genetics **118:** 353-363.

HUNTER, N., and N. KLECKNER, 2001 The single-end invasion: an asymmetric intermediate at the double-strand break to double-holliday junction transition of meiotic recombination. Cell **106:** 59-70.

KIMURA, M., and T. OHTA, 1969 The average number of generations until fixation of a mutant gene in a finite population. Genetics **61:** 763-771.

LAUGHNAN, J. R., 1953 The effect of the *sh2* factor on carbohydrate reserves in the mature endosperm of maize. Genetics **38:** 485-499.

LICHTEN, M., and A. S. GOLDMAN, 1995 Meiotic recombination hotspots. Annu. Rev. Genet. **29:** 423-444.

MAINS, E. B., 1949 Heritable characters in maize. Linkage of a factor for shrunken endosperm with the *a1* factor for aleurone color. J. Hered. **40:** 21-24.

MATSUOKA, Y., Y. VIGOUROUX, M. M. GOODMAN, G. J. SANCHEZ, E. BUCKLER and J. DOEBLEY, 2002 A single domestication for maize shown by multilocus microsatellite genotyping. Proc. Natl. Acad. Sci. USA. **99:** 6080-6084.

MODRICH P., and R. LAHUE, 1996 Mismatch repair in replication fidelity, genetic recombination, and cancer biology. Annu. Rev. Biochem. **65:** 101-133.

PAQUES, F., and J. E. HABER, 1999 Multiple pathways of recombination induced by double-strand breaks in Saccharomyces cerevisiae. Microbiol Mol. Biol. Rev. **63:** 349-404.

PENNISI, E., 2004 The biology of genomes meeting: The case of the disappearing DNA hotspots. Science **304:** 1590.

PETES, T. D., 2001 Meiotic recombination hot spots and cold spots. Nat. Rev. Genet. **2:** 360-369.

PHILLIPS, R. L., 1969 Recombination in *Zea mays* L. 11. Cytogenetic studies of recombination in reciprocal crosses. Genetics **61:** 117-127.

PUCHTA, H., and B. HOHN, 1996 From centiMorgans to base pairs: homologous recombination in plants. Trends Genet. **1:** 340-348.

RAFALSKI A., and M. MORGANTE, 2004 Corn and humans: recombination and linkage disequilibrium in two genomes of similar size. Trends Genet. **20:** 103-111.

RHOADES, M. M., 1978 Genetic effects of heterochromatin in maize, pp. 641-671 in *Maize Breeding and Genetics*, edited by D. B. WALDEN. Wiley & Sons, New York.

ROBERTSON, D. S., 1967 Crossing over and chromosomal segregation involving the B^9 element of the A-B translocation B-9b in maize. Genetics **55:** 433-449.

ROBERTSON, D. S., 1984 Different frequency in the recovery of crossover products from male and female gametes of plants hypoploid for B-A translocations in maize. Genetics **107:** 117-130.

SCHNABLE, P. S., A. P. HSIA and B. J. NIKOLAU, 1998 Genetic recombination in plants. Curr. Opin. Plant Biol. **1:** 123-129.

SUN, H., D. TRECO, and J. W. Szostak, 1991 Extensive 3'-overhanging, single-stranded DNA associated with the meiosis-specific double-strand breaks at the ARG4 recombination initiation site. Cell **64:** 1155-1161.

SZOSTAK, J. W., T. L. ORR-WEAVER, R. J. ROTHSTEIN and F. W. STAHL, 1983 The double-strand-break repair model for recombination. Cell **33:** 25-35.

TIMMERMANS, M. C., O. P. DAS, J. M. BRADEEN and J. MESSING, 1997 Region-specific cis- and trans-acting factors contribute to genetic variability in meiotic recombination in maize. Genetics **146:** 1101-1113.

THURIAUX, P., 1977 Is recombination confined to structural genes on the eukaryotic genome? Nature **268:** 460-462.

TUERCK, J. A., and M. E. FROMM, 1994 Elements of the maize A1 promoter required for transactivation by the anthocyanin B/C1 or phlobaphene P regulatory genes. Plant Cell **6:** 1655-1663.

WANG, R. L., A. STEC, J. HEY, L. LUKENS and J. DOEBLEY, 1999 The limits of selection during maize domestication. Nature **398:** 236-239.

XU, X., A. P. HSIA, L. ZHANG, B. J. NIKOLAU and P. S. SCHNABLE, 1995 Meiotic recombination break points resolve at high rates at the 5' end of a maize coding sequence. Plant Cell **7:** 2151-2161.

YAO, H., Q. ZHOU, J. LI, H. SMITH, M. YANDEAU, B. J. NIKOLAU and P. S. SCHNABLE, 2002 Molecular characterization of meiotic recombination across the 140-kb multigenic *a1-sh2* interval of maize. Proc. Natl. Acad. Sci. USA. **99:** 6157-6162.

Figure legends

Figure 1. Comparisons of recombination rates/Mb and distributions of recombination breakpoints among stocks that carry different *A1 Sh2* haplotypes. Rates and distributions were compared via χ^2 tests and p values are indicated. Statistically significant differences are indicated by asterisks. *: Significant difference at the 0.05 level; **: Significant

difference at the 0.01 level. The LC haplotype in panel A is carried in the LC2 stock; the LC haplotype in panel C is carried in the LC1 stock. Comparisons in panel D did not include recombinants that resolved in subinterval V-1 (Figure 4) because the sizes of this subinterval vary too much among haplotypes to permit fair comparisons. The distribution of breakpoints across the *yz1* gene considered only the transcribed region (subinterval IV).

Figure 2. Distributions of recombination events and sequence polymorphisms across the *a1-sh2* intervals of the mex, par and lux haplotypes. (A) Comparisons of the structures of the three teosinte *A1 Sh2* haplotypes relative to the maize *a1::rdt sh2* haplotype. Genes are indicated as boxes. The polymorphisms relative to the *a1::rdt sh2* haplotype used to define the subintervals between the *a1* and *sh2* loci are indicated by gray dashed lines. Subintervals I, II, IV and VI were completely sequenced for all haplotypes. Subinterval III was completely sequenced for the mex, lux and *a1::rdt sh2* haplotypes, and partially sequenced for the par haplotype. Large InDels in subinterval III are indicated by triangles (insertions) and parentheses (deletions). The *rdt* transposon insertion was indicated by a triangle. Large Indels in other subintervals were not shown. Haplotype-specific IDP primers used to map recombination breakpoints are indicated by horizontal arrows. The sizes of each subinterval are based on those of the *a1::rdt sh2* haplotype. (B) Observed percentages of recombinants that resolved in each subinterval. The "*" and "**" indicate significant differences between the rates of recombination/Mb based on the observed recombination breakpoints mapped to subintervals and the corresponding average rates across the *a1-sh2* interval of each haplotypes at the 0.05 and 0.01 levels, respectively. (C) Percentages of recombinants expected to resolve in each subinterval based on a random distribution across the *a1-sh2*

interval. (D) Recombination rates/Mb in subintervals. The indicated average rates of recombination/Mb across the *a1-sh2* interval in each of the three stocks were calculated based on the physical size (~130 kb) of the common *a1::rdt sh2* haplotype carried in all stocks. The horizontal dashed line indicates the average recombination rate/Mb of the maize genome (2.1 cM/Mb). (E) Frequencies of sequence polymorphisms (#/100 bp) between each *A1 Sh2* haplotype and the *a1::rdt sh2* haplotype. Numbers of SNPs /InDels in each of the subintervals are presented. These numbers for subinterval III-par were calculated from the sequenced portions of this subinterval. ND: not determined due to incomplete sequences in the indicated subintervals.

Figure 3. High-resolution mapping of the recombination breakpoints associated with the LC, mex, par and lux haplotypes that resolved in the *a1* locus. (A) – (D) Exons of the *a1* gene are shown as boxes. Short vertical lines represent sequence polymorphisms between *A1* haplotypes and the *a1::rdt* haplotype. The widths of the vertical lines are proportional to the numbers of polymorphic nucleotides. Subintervals are defined by sequence polymorphisms. Haplotype-specific primers are indicated by horizontal arrows. The numbers of recombination breakpoints that mapped to each subinterval for each haplotype are shown. Large InDels are indicated by triangles (insertions) and parentheses (deletions). Panel A depicts the positions of recombination breakpoints previously characterized by YAO *et al.* 2002, but here classified relative to subintervals I-1 and I-2. (E) Comparison of recombination rates/Mb across the *a1* locus among the LC, mex, par and lux haplotypes. The horizontal dashed line indicates the average recombination rate/Mb of the maize genome (2.1 cM/Mb). The "**" indicates that the recombination rate/Mb in the labeled haplotype at the

corresponding subinterval is significantly different from all others at the 0.01 level. (F) Comparison of frequencies of sequence polymorphisms (#/100 bp) at the *a1* locus among the LC, mex, par and lux haplotypes. Sequence polymorphisms are between each *A1 Sh2* haplotype and the *a1::rdt sh2* haplotype. Numbers of SNPs/InDels in each of the subintervals are also listed.

Figure 4. High-resolution mapping of the recombination breakpoints associated with the mex, par and lux haplotypes that resolved in the *yz1* locus. (A) – (C) Exons of the *yz1* gene are shown as boxes. Short vertical lines represent sequence polymorphisms between each teosinte *Yz1* allele and the *Yz1* allele from the *a1::rdt sh2* stock. The widths of these short vertical lines are proportional to the numbers of polymorphic nucleotides. Subintervals are defined by sequence polymorphisms. Haplotype-specific primers are indicated by horizontal arrows. The numbers of recombination breakpoints that mapped to each subinterval are shown for each haplotype. Large InDels are indicated by triangles (insertions) and parentheses (deletions). (D) Comparison of recombination rates/Mb across the *yz1* locus among the mex, par and lux haplotypes. The horizontal dashed line indicates the average recombination rate/Mb of the maize genome (2.1 cM/Mb). The "**" indicates that the recombination rate/Mb in the labeled haplotype at the corresponding subinterval is significantly different from the others at the 0.01 level. (E) Comparison of the frequencies of sequence polymorphisms (#/100 bp) at the *yz1* locus among the mex, par and lux haplotypes. Numbers of sequence polymorphisms were calculated by comparing each of the teosinte *Yz1* alleles and the common *Yz1* allele from the *a1::rtdt sh2* stock. Numbers of SNPs/ InDels in each of the subintervals are listed.

Figure 5. Recombination breakpoints across the *a1*-Interloop region in the mex haplotype. Exons of the *a1* gene are shown as boxes. Short vertical lines represent sequence polymorphisms between the mex *A1 Sh2* haplotype and the *a1::rdt sh2* haplotype. The widths of the vertical lines are proportional to the numbers of polymorphic nucleotides. Subintervals are defined by sequence polymorphisms. Haplotype-specific primers are indicated by horizontal arrows. The numbers of recombination breakpoints that mapped to each subinterval are shown. Large InDels are indicated by triangles.

TABLE 1

Oligonucleotides used as primers for PCR and sequencing

Primer Name	Sequences[a]	Haplotypes[b]			
		a1::rdt sh2	mex	par	lux
rdt444	AGCAAATAGCAATAATCAAGGCA	+[c]	-[d]	-	-
aIDPrdt4	AATTAGTCTCTCGATCATCT	+	-	-	-
aIDPrdt3	CTAAAGAAGCAAAGCAA	+	-	-	-
yzIDPrt5	GCATGTTAAAAATAGAAGAAG	+	-	-	-
yzIDPrt4	TTCACACAAAAAAAGGC	+	-	-	-
yzIDPrt3	CTAGGAGTACATGTTTTTTC	+	-	-	-
IDPrdtx	TAATTCTAGTGTCCCAAC	+	-	-	-
QZ1001	GATACAGAAGTATATATAAGGGCCAA	+	+	-	-
a1rdt2912	AACACCCCGCTAACAC	+	+	-	-
a1rdt1541	CGCTAACTATCTCGGTAACT	+	+	-	-
QZ1002	TATTCGTAATGATGTTTAT	+	-	+	-
ajl001	GGAGAGTCGAATAAAAAGTGT	+	+	+	-
a1rdt2381	TCAACCGTGCTACCAACT	+	+	+	-
IrlL3	ATCGGCAAACCCACCAA	+	+	-	+
ZH792	GCGGTTGCGGCTTGT	+	+	-	+
IDPIRmex	GTAAGTCTCTATCCAGTC	-	+	-	-
YZ4725	AAATGGTCAGGATAGCTTAGTT	-	+	-	-
ZH1384	GCCATCTCTACTGTTACCTT	-	+	-	-
IDPyz5lr	TATCAAGCACAAGCAG	-	-	-	+
yzIDPmpl	AGTAGAGAGGAAATCAGAAG	-	+	+	+
A1.2	GATTGTTGCTTAAGCGCCAATCGT	+	+	+	+
AE4EI	CGAATTCCGCCAGGGTTTTAGACA	+	+	+	+
XX390	TCGGCTTGATTACCTCATTCT	+	+	+	+
yz3utrf	CGGGGGTTGCAGTCATTGAC	+	+	+	+
YZ3	GGAAGCCTGTTTTGGTG	+	+	+	+
yz4127F	CATCATCTCCGTGTTCTC	+	+	+	+
ZH1748	CACATCCCCGTCTCCT	+	+	+	+
ZH2617	CGAACAGGGAAGAATGG	+	+	+	+
YZ1	GCGGCGTTGCTGCTGTA	+	+	+	+
YZc85	GGAGACGGGGATGTGG	+	+	+	+
XL2	TGTTCAAAGTGGGAGG	+	+	+	+

[a] Sequences are listed 5' to 3'.
[b] The mex, par and lux haplotypes are *A1 Sh2*.
[c] + indicates a primer that can amplify the corresponding haplotype.
[d] – indicates a primer that cannot amplify the corresponding haplotype.

TABLE 2

Isolation of recombinants from stocks carrying distinct *A1 Sh2* haplotypes

Stocks	*A1 Sh2*	Year	No. isolated			No. tested[b]			No. confirmed			No. corrected[c]			Popl. size	Genetic distance (cM)[d]
			Clsh[a]	clrd[a]	Total	Clsh	clrd	Total	Clsh	clrd	Total	Clsh	clrd	Total		
mex	mex	1997	140	145	285	80	106	186	76	101	177	133	138	271	133,040	0.20 ± 0.012
par	par	1997	34	52	86	22	40	62	22	39	61	34	51	85	87,515	
		1998	59	77	136	15	34	49	15	30	45	59	68	127	116,838	
		Pooled	93	129	222	37	74	111	37	69	106	93	120	213	204,353	0.10 ± 0.0071
lux	lux	1997	83	83	166	62	68	130	60	58	118	80[f]	71[f]	151[f]	2,07,184	
		1998[e]	60	144	204	12	59	71	11	56	67	55[e,f]	137[e,f]	192[f]	319,622	
		Pooled	143	227	370	74	127	201	71	114	185	137	204	341	526,806	0.065 ± 0.0035
LC2	LC	1998	13	13	26	2	11	13	2	11	13	13	13	26	27,868	0.093 ± 0.018

[a] Clsh, colored shrunken kernels; clrd, colorless round kernels.

[b] Putative recombinants were tested by genetic crosses and/or molecular analysis, *e.g.*, PCR mapping of recombination breakpoints as described by YAO *et al.* (2002).

[c] No. corrected = No. isolated x (No. confirmed / No. tested).

[d] Calculated as No. corrected / Population size x 100. See also methods.

[e] The ratio of No. corrected Clsh to clrd is significantly different from 1:1 (p-value = $3.6 \times e^{-9}$).

[f] Although the corrected numbers of Clsh are significantly different between 1997 and 1998 (p-value = $3.3 \times e^{-6}$), the corrected numbers of clrd and the corrected numbers of total recombinants are not.

TABLE 3

Statistical analyses of recombination in the seven subintervals of the *a1-sh2* interval

Subintervals	Haplotypes	Comparisons to the average of *a1-sh2*[a]	Comparisons to the genome's average[b]	Comparisons among stocks[c]	Features[d]
I	mex	0.28	0.40	0.016↓ (mex *vs.* par)	average spot
	par	$6.7e^{-5}$↑	0.00032↑	0.0014↑ (par *vs.* lux)	**hot** spot (local, global)
	lux	0.00014↑	0.010↑	0.87 (lux *vs.* mex)	**hot** spot (local, global)
II	mex	$1.3e^{-8}$↑	$2.1e^{-8}$↑	$9.9e^{-8}$↑ (mex *vs.* par)	**hot** spot (local, global)
	par	0.14	0.28	0.11 (par *vs.* lux)	average spot
	lux	0.34	0.80	$<2.2e^{-16}$↓ (lux *vs.* mex)	average spot
III	mex	0.00033↑	0.0019↑	$9.3e^{-6}$↑ (mex *vs.* par)	**hot** spot (local, global)
	par	0.99	0.25	0.0013 (par *vs.* lux)	average spot
	lux	0.026↓	$1.9e^{-7}$↓	$<2.2e^{-16}$↓ (lux *vs.* mex)	**cold** spot (local, global)
IV	mex	$<2.2e^{-16}$↑	$<2.2e^{-16}$↑	0.011↑ (mex *vs.* par)	**hot** spot (local, global)
	par	$2.9e^{-14}$↑	$5.5e^{-12}$↑	0.026↑ (par *vs.* lux)	**hot** spot (local, global)
	lux	$<2.2e^{-16}$↑	$<2.2e^{-16}$↑	$1.9e^{-8}$↓ (lux *vs.* mex)	**hot** spot (local, global)
V	mex	$1.2e^{-9}$↓	$2.9e^{-13}$↓	0.67 (mex *vs.* par)	**cold** spot (local, global)
	par	0.00028↓	$4.6e^{-14}$↓	0.58 (par *vs.* lux)	**cold** spot (local, global)
	lux	$1.9e^{-6}$↓	$<2.2e^{-16}$↓	0.89 (lux *vs.* mex)	**cold** spot (local, global)
VI	mex	0.049↑	0.083	0.037↑ (mex *vs.* par)	**hot** spot (local)
	par	0.99	0.85	0.55 (par *vs.* lux)	average spot
	lux	0.020↑	0.50	0.044↓ (lux *vs.* mex)	**hot** spot (local)
VII	mex	$<2.2e^{-16}$↓	$<2.2e^{-16}$↓	0.083 (mex *vs.* par)	**cold** spot (local, global)
	par	$1.3e^{-15}$↓	$<2.2e^{-16}$↓	0.94 (par *vs.* lux)	**cold** spot (local, global)
	lux	$<2.2e^{-16}$↓	$<2.2e^{-16}$↓	0.037↓ (lux *vs.* mex)	**cold** spot (local, global)

TABLE 3 (continued)

[a-c] Goodness-of-fit χ^2 tests were used in the comparisons of the observed rate of recombination in a given subinterval to the average rate of recombination in each teosinte *A1 Sh2* haplotype (a), to the genome's average (2.1 cM/Mb) (b) and the comparisons of rates of recombination in a given subinterval among the three teosinte *A1 Sh2* haplotypes (c). Details were described in the Methods. The p values obtained from these χ^2 tests are listed. The ↑ and ↓ indicates that an observed rate of recombination is significantly higher and lower (at the 0.05 level), respectively, than the rate of recombination to which it was compared.

[d] According to its recombination activity, a subinterval is classified as a global or local hot spot, an average spot or a global or local cold spot. A global hot or cold spot exhibits significantly higher or lower recombination activity than the genome as a whole. A local hot or cold spot exhibits significantly higher or lower recombination activity than the *a1-sh2* interval. Recombination activity of an average spot is not significantly different from those of the *a1-sh2* interval and the genome. The cutoff level for the p values is 0.05.

TABLE 4

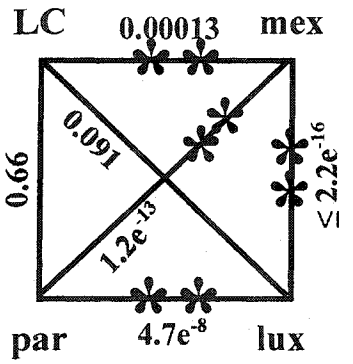
Distributions of recombination breakpoints across the *a1* and *yz1* loci

Haplotypes	*a1* locus [a]	*yz1* locus [a]
mex	$1.9e^{-5}$	0.00090
par	0.0024	0.00037
lux	0.20	$4.1e^{-9}$
LC	0.032	ND [b]

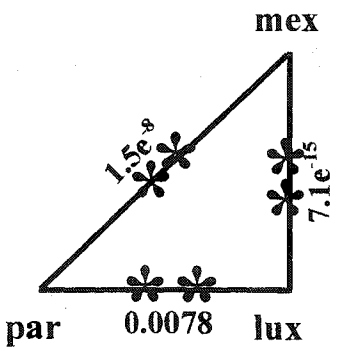
[a] The observed distributions of recombination breakpoints across the *a1* (Figure 3, subintervals I–II) and *yz1* (Figure 4, subintervals IV–V-1) loci in each *A1 Sh2* haplotypes were compared to the expected distributions under the assumption of a random distribution across the *a1* and *yz1* using the χ^2 contingency tests. The p values from these tests are shown.

[b] ND, not done.

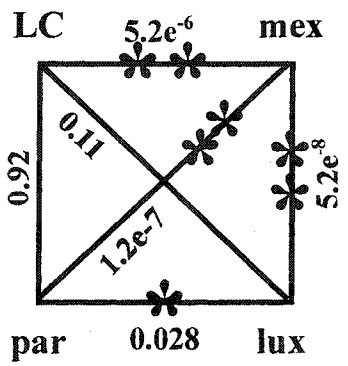
A. Rates of recombination/Mb between the *a1* and *sh2* loci.



B. Distributions of recombination breakpoints across the *a1-sh2* interval.



C. Distributions of recombination breakpoints across the *a1* locus (subintervals I-II).



D. Distributions of recombination breakpoints across the *yz1* locus (subinterval IV).

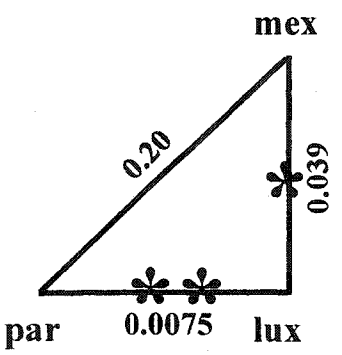
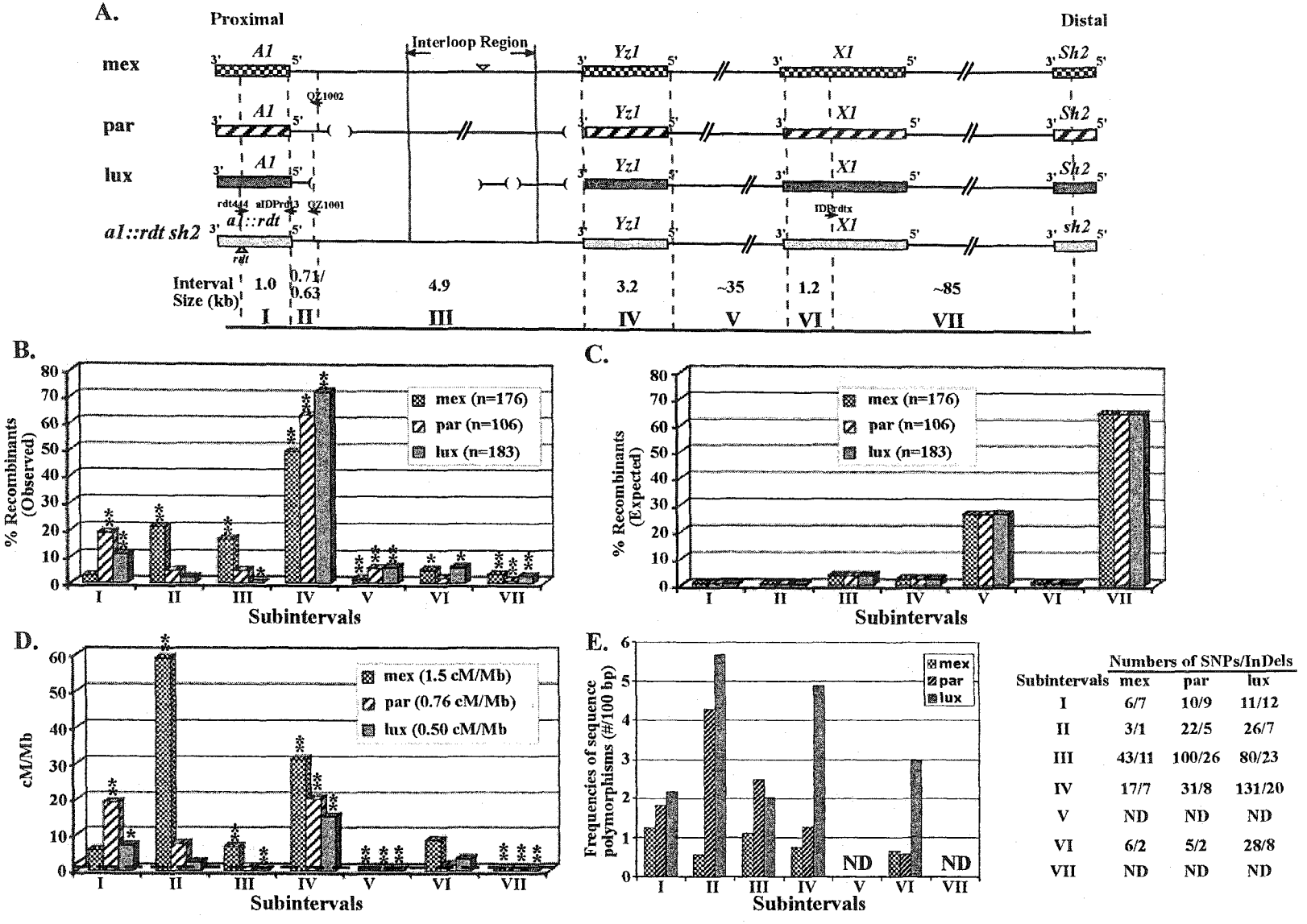


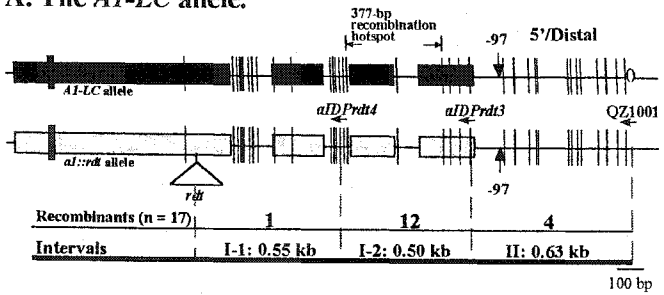
Figure 1. YAO *et al.*



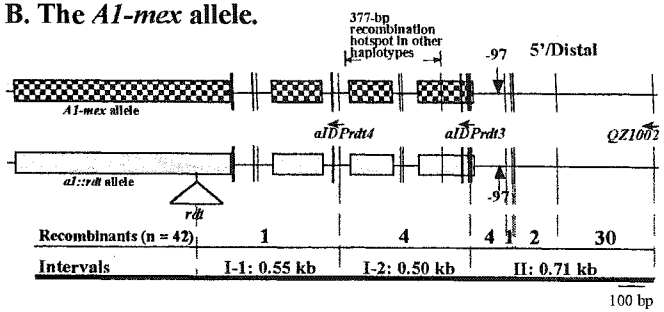
Subintervals	Numbers of SNPs/InDels mex	par	lux
I	6/7	10/9	11/12
II	3/1	22/5	26/7
III	43/11	100/26	80/23
IV	17/7	31/8	131/20
V	ND	ND	ND
VI	6/2	5/2	28/8
VII	ND	ND	ND

Figure 2. YAO *et al.*

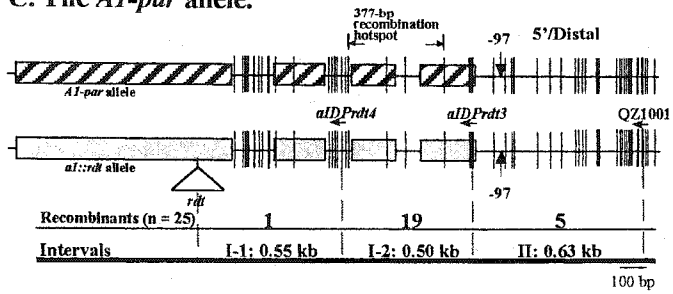
A. The *A1-LC* allele.



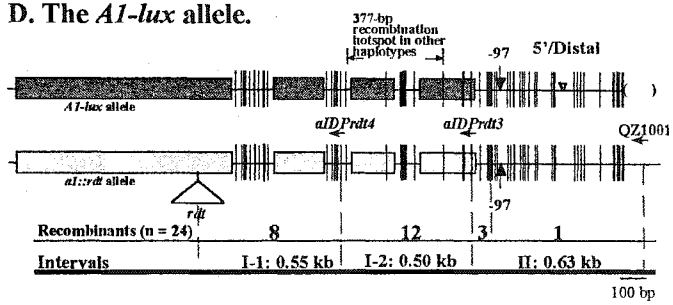
B. The *A1-mex* allele.



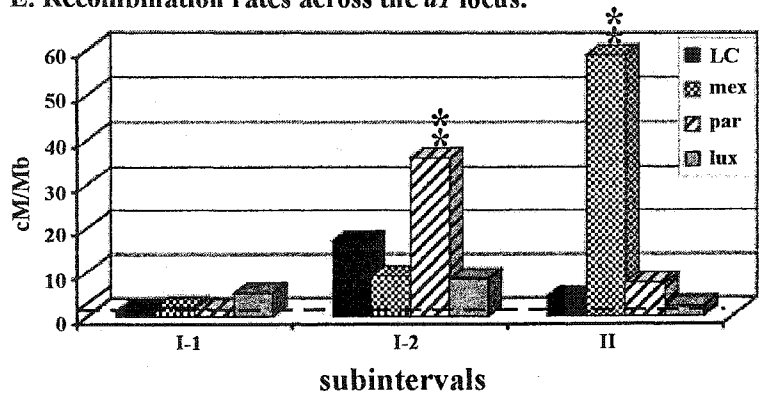
C. The *A1-par* allele.



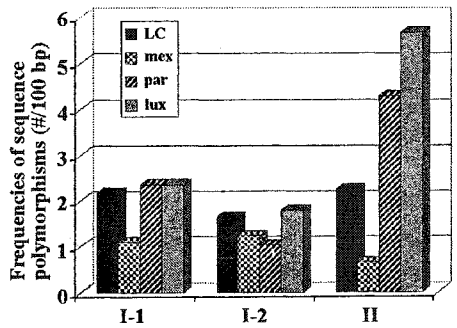
D. The *A1-lux* allele.



E. Recombination rates across the *a1* locus.

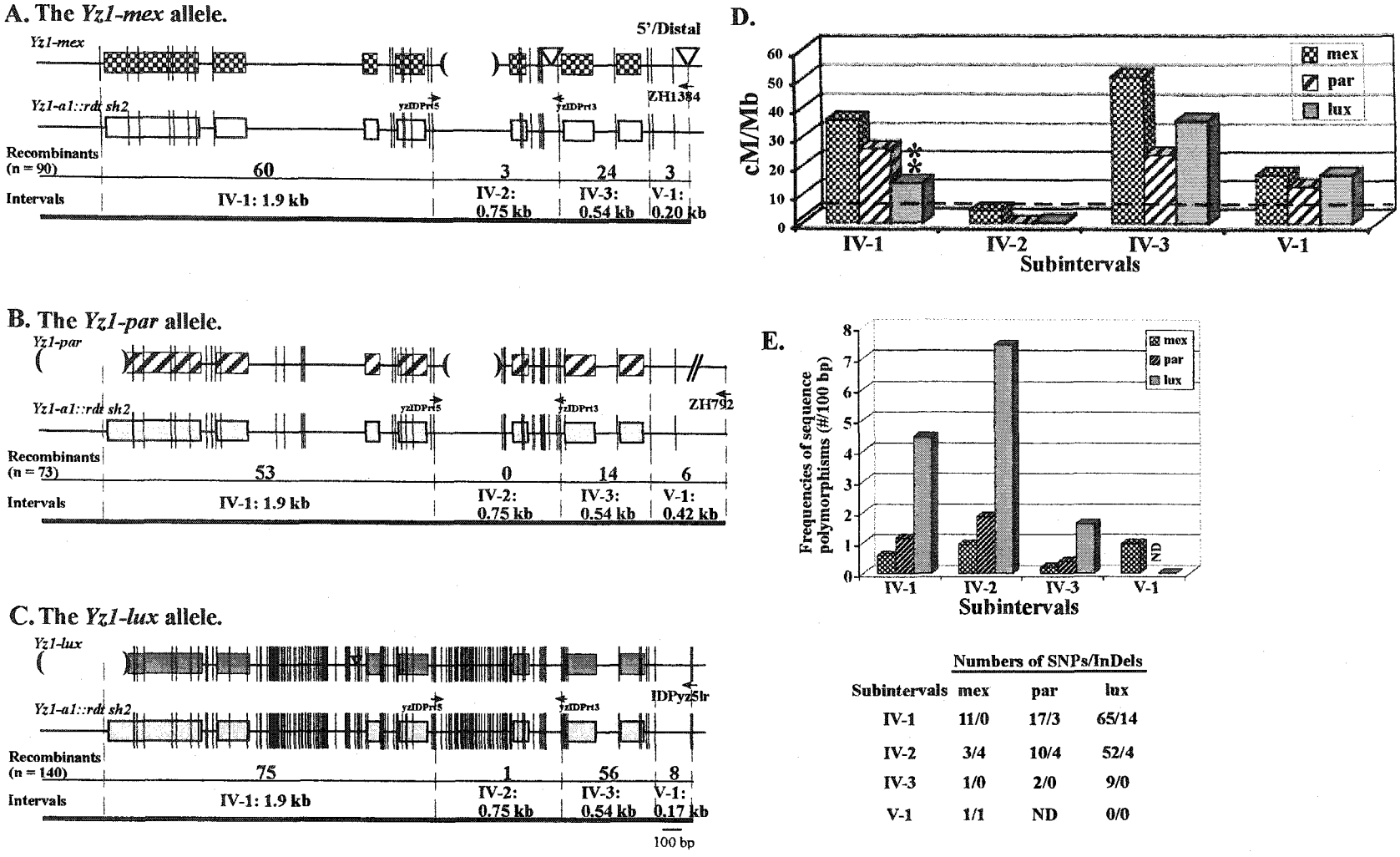


F. Sequence comparisons across the *a1* locus.



	Numbers of SNPs/InDels			
subintervals	LC	mex	par	lux
I-1	7/5	2/4	7/6	6/7
I-2	6/2	4/2	3/2	5/4
II	11/3	3/1	22/5	26/7

Figure 3. YAO *et al.*



	Numbers of SNPs/InDels		
Subintervals	mex	par	lux
IV-1	11/0	17/3	65/14
IV-2	3/4	10/4	52/4
IV-3	1/0	2/0	9/0
V-1	1/1	ND	0/0

Figure 4. YAO *et al.*

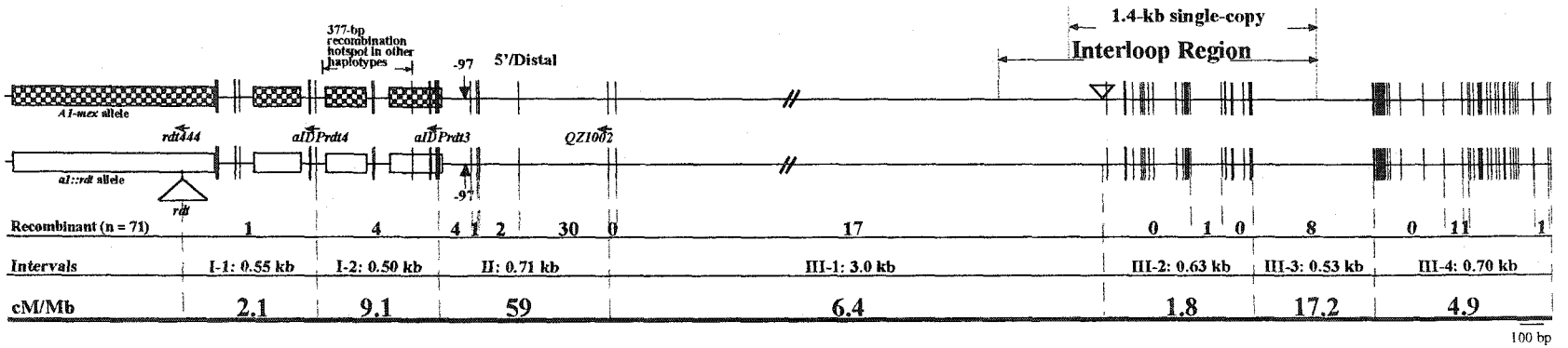


Figure 5. YAO *et al.*

CHAPTER 4. EVALUATION OF FIVE *AB INITIO* GENE PREDICTION PROGRAMS FOR THE DISCOVERY OF MAIZE GENES

A paper to be submitted to *Plant Molecular Biology*

Hong Yao, Ling Guo[1], Yan Fu[1], Lisa A. Borsuk, Tsui-Jung Wen, David S. Skibbe, Xiangqin Cui[2], Brian E. Scheffler, Jun Cao, Scott J. Emrich, Daniel A. Ashlock and Patrick S. Schnable

Abstract

Five *ab initio* programs (FGENESH, GeneMark.hmm, GENSCAN, GlimmerR and Grail) were evaluated for their accuracy in predicting maize genes. Two of these programs, GeneMark.hmm and GENSCAN had been trained for maize; FGENESH had been trained for monocots (including maize), and the others had been trained for rice or *Arabidopsis*. Initial evaluations were conducted using eight maize genes (*gl8a*, *pdc2*, *pdc3*, *rf2c*, *rf2d*, *rf2e1*, *rth1*, and *rth3*) the sequences of which were not released to the public prior to conducting this evaluation. The significant advantage of this data set for this evaluation is that these genes could not have been included in the training sets of the prediction programs. FGENESH yielded the most accurate and GeneMark.hmm the second most accurate predictions. The five programs were used in conjunction with RT-PCR to identify and establish the structures of two new genes in the *a1-sh2* interval of the maize genome. FGENESH, GeneMark.hmm and GENSCAN were tested on a larger data set consisting of maize assembled genomic islands (MAGIs) that had been aligned to ESTs. FGENESH, GeneMark.hmm and GENSCAN correctly predicted gene models in 773, 625, and 371 MAGIs, respectively, out of the 1,353 MAGIs that comprise data set 2.

---

[1] These authors contributed equally to this report.

Introduction

Locating the positions of all the genes and determining their structures is a first step toward deciphering the functions of a sequenced genome. Two approaches are available (reviewed by Stormo, 2000; Pertea and Salzberg, 2002; Mathé *et al.*, 2002). The first is based on sequence similarity. A significant degree of sequence identity or similarity between a genomic query sequence and cDNA, EST, protein or genomic sequences of a gene from the same or another species can provide evidence that a query sequence contains a gene. This method is, however, highly dependent upon the quantity and quality of pre-existing sequence data. Typically only 50 to 70% of the genes in a sequenced genome can be found via comparisons to other genomes, although this fraction will increase as the number of sequenced genomes increases (reviewed by Pertea and Salzberg, 2002; Mathé *et al.*, 2002). In addition, sequence similarity searches can provide misleading information due to artifacts in databases. The second approach for identifying genes in a sequenced genome is to use *ab initio* gene prediction programs. *Ab initio* gene prediction uses statistical and computational methods to detect coding regions, splice sites, and start and stop codons in genomic sequences. This approach does not depend on sequence similarity and is therefore not limited by the availability of sequence data. But as compared to predictions based on sequence similarity, *ab initio* predictions are currently typically less accurate because available programs are not yet able to make highly reliable predictions of gene structures. One reason for this is that the quality of predictions is limited by the quality of the training sets. These training sets usually consist of gene sequences that have been characterized in a given species.

To date only two plant genomes, *Arabidopsis* (The *Arabidopsis* Genome Initiative, 2000) and rice (Yu *et al.*, 2002; Goff *et al.*, 2002), have been completely sequenced. Efforts to sequence other crop genomes, including maize, are underway (http://www.nsf.gov/bio/pubs/awards/genome02.htm) (Palmer *et al.*, 2003; Whitelaw *et al.*, 2003). The maize genome consists of about 2,400 Mb, i.e., approximately 6-fold larger than that of rice (reviewed by Moore, 2000). It is estimated that the maize genome contains approximately 50,000 genes that account for only 10-15% of the genome (Bennetzen *et al.*, 2001). Much of the genome is repetitive elements, many of which are retrotransposons (SanMiguel *et al.*, 1996). Due to the large size and highly repetitive nature of the maize genome, sequencing efforts are being focused on the gene-rich, low-copy fraction of the genome, i.e., the "gene space". Two methods are being used to isolate the "gene space", methylation-filtration (MF) (Rabinowicz *et al.*, 1999) and high C_0t (HC) selection (Peterson *et al.*, 2002; Yuan *et al.*, 2003).

The identification of genes from sequences generated from the maize genome sequencing project will establish whether the current sequencing approaches are successfully enriching for genes, and will, in addition, define genomic resources necessary to study the function of maize. Given the limitations associated with gene prediction based on sequence similarity, *ab initio* gene prediction programs will necessarily play an important role in maize gene discovery. In an effort to develop an *ab initio* gene discovery strategy for maize, existing versions of five programs (Table 1) including FGENESH (Salamov and Solovyev, 2000), GeneMark.hmm (Lukashin and Borodovsky, 1998), GENSCAN (version 1.0) (Burge and Karlin, 1997), GlimmerR (Salzberg *et al.*, 1999; Yuan *et al.*, 2001) and Grail (version 1.3) (Xu and Uberbacher, 1997) were evaluated for their accuracy in predicting maize genes.

The purpose of this study was to evaluate currently available tools for suitability in the *ab initio* discovery of genes from partial maize genomic sequence rather than to compare the algorithms that underlie these tools. Maize-trained versions of three of these programs, FGENESH, GeneMark.hmm and GENSCAN, are available. For the remaining programs versions that had been trained using rice (GlimmerR) or *Arabidopsis* (Grail) were used. Each program was evaluated using a data set consisting of genomic sequences of eight genes cloned by us (Table 2). Because these gene sequences could not have been included in the data sets used to train the *ab initio* programs, they represent a valuable tool for evaluating these programs. These five programs were also used to help identify and determine the structures of two genes in the 140-kb maize *a1-sh2* interval (Civardi *et al.*, 1994; Yao *et al.*, 2002). FGENESH, followed by GeneMark.hmm and GENSCAN made more accurate gene predictions in these tests. Their ability to predict maize genes was further tested using a larger data set (1,353 genic sequences) consisting of maize assembled genomic islands (MAGIs) assembled from genome survey sequences (GSSs) whose exons were identified via alignments to ESTs. In this larger data set, FGENESH was still the most accurate program.

Materials and methods

Data Sets

Two data sets were used to evaluate the accuracy of gene prediction programs.

Data set 1: The genomic sequences of eight maize genes (*gl8a*, *pdc2*, *pdc3*, *rf2c*, *rf2d*, *rf2e1*, *rth1*, and *rth3*) that were cloned in our lab but that had not been released to the public prior to the completion of this evaluation constitute data set 1. With the exception of *rf2d* that is incomplete at its 5' end, all of these genic sequences contain the corresponding start

and stop codons. The GenBank accession numbers of each gene sequence are listed in Table 2. Their gene structures were determined by spliced alignment of full-length cDNA sequences to the corresponding genomic sequences using the GeneSeqer program (http://bioinformatics.iastate.edu/cgi-bin/gs.cgi) (Usuka *et al.*, 2000; Usuka and Brendel, 2000, Brendel *et al.*, 2004). To make a fair comparison of the predictions among genes in this data set, genomic sequences that contain complete genes were trimmed at their 5' and 3' ends. The incomplete *rf2d* genomic sequence was only trimmed at the 3' end. Consequently, in this data set, the amount of sequence before the start codon of each gene is 520 bp and after the stop codon is 375 bp. The statistical characteristics of each gene in data set 1 are listed in Table 2.

Data set 2: Data set 2 consists of a subset of the 114,173 ISU MAGIs in version 3.1b (http://plantgenomics.iastate.edu/maize). These MAGIs were assembled from 879,523 GSSs (MF and HC sequences) of the maize inbred line B73 using a strategy similar to that described by Emrich *et al.* (2004) (see Supplementary Materials). MAGIs to include in Data set 2 were selected based on the qualities of their GeneSeqer alignments to clustered B73 ESTs generated by Schnable Lab. Detailed methods used to generate data set 2 are provided in the Supplementary Materials. In summary, data set 2 consists of 1,353 selected MAGI contigs that contain at least one pair of reliable donor and acceptor sites flanking an intact intron (Figure 3). Alignments between the selected genomic sequences in data set 2 and the corresponding EST sequences are available from the authors upon request. The statistical characteristics of data set 2 are shown in Figure 1.

Statistical comparison of data set 1 to 74 structure-known genes of maize

The GC contents and lengths of internal exons and introns in data set 1 were compared to those in a data set consisting of 74 structure-known maize genes. The GC contents and lengths of the exons and introns of the 74 structure-known maize genes were calculated by parsing the exons and introns from the sequences that were downloaded from NCBI. Only complete internal exons and introns were used in this analysis. The length of each internal exon and intron was determined and then the number of G's and C's were counted and divided by the total length to determine the percent GC content. Similar calculations were conducted for sequences in data set 1. The two-sample Kolmogorov-Smirnov test (Kolmogorov, 1933; Smirnov, 1939), which tests the null hypothesis that the data values from two samples have the same continuous distribution, was used to compare these parameters.

Programs evaluated

Five *ab initio* programs were evaluated using data set 1. The features of these programs are listed in Table 1. Available versions of FGENESH, GeneMark.hmm and GENSCAN (version 1.0) that had been trained for maize were evaluated. For those two programs for which a maize trained version was not available, the version trained for the closest organism to maize was evaluated. Although all five programs make predictions in both strands of a genomic DNA sequence, only the predictions for the strand containing known genic sequences were analyzed in this study because they could be compared with the known actual gene structures or splice sites in our test data set 1 and 2. FGENESH, GeneMark.hmm, GENSCAN and GlimmerR predict gene models that can be single or multiple in a genomic sequence and a predicted exon is indicted as initial (starting with the

initiation codon and ending with a donor site), internal (starting with an acceptor site and ending with a donor site), terminal (starting with an acceptor site and ending with the stop codon) or single (starting with the initiation codon and ending with the stop codon) exon in the output. Grail predicts a series of non-overlapping exons in both DNA strands but no gene model is produced. All programs were run via their web sites by using their organism-specific default parameters to obtain the prediction results for data set 1 (Table 1). To obtain predictions for data set 2, FGENESH, GeneMark.hmm and GENSCAN were run locally using the default parameters for monocot sequences (including maize) with usage of the GC donor site (FGENESH) or using the default parameters for maize sequences (GeneMark.hmm and GENSCAN). Additional information is provided in the Supplementary Materials. SplicePredictor (Brendel *et al.*, 2004), NetGene2 (Hebsgaard *et al.*, 1996; Tolstrup *et al.*, 1997) and GeneSplicer (Pertea *et al.*, 2001) were not evaluated in this study because they do not predict gene models.

Evaluation of gene prediction programs

The performance of each program was evaluated at three levels (splice site, nucleotide and exon) as described by Burset and Guigo (1996) and Pavy *et al.* (1999).

At the splice site level, the accuracy of a program's predictions is measured by SN, SP and the average of SN and SP ((SN + SP)/2). If true positive (TP) is defined as the number of correctly predicted splice sites, FP as the number of incorrectly predicted splice sites, and false negative (FN) as the number of actual splice sites missed in the prediction, then SN = TP/(TP + FN) and SP = TP/(TP + FP). Since neither SN nor SP alone can

represent the accuracy of a program, the value of (SN + SP)/2 is usually used as a measure of accuracy.

SN and SP are also used to evaluate predictions at the nucleotide level. Here TP is the number of nucleotides that are correctly predicted as coding, TN is the number of nucleotides that are correctly predicted as non-coding, FP is the number of nucleotides that are incorrectly predicted as coding, and FN is the number of nucleotides that are incorrectly predicted as non-coding. Under these definitions, SN = TP/(TP + FN), SP = TP/(TP + FP). The value of the Correlation Coefficient (CC) that reflects both SN and SP is used for evaluation. CC is defined as:

$CC = ((TP \times TN) - (FN \times FP))/((TP + FN) \times (TN + FP) \times (TP + FP) \times (TN + FN))^{1/2}$.

At the exon level, if AE is defined as the actual exons, TE as the correctly predicted exons, , PE as the predicted exons that are partially correct (i.e., only one boundary correct), OE as the predicted exons that overlap with the actual exons but with both boundaries wrong, ME as the actual exons missed in the prediction, WE as the number of incorrectly predicted exons, then SN = TE/AE, SP = TE/(TE + PE + OE + WE), PE% = 100 x PE/(TE + PE + OE + WE), OE% = 100 x OE/(TE + PE + OE + WE), ME% = 100 x ME/AE, and WE% = 100 x WE/(TE + PE + OE + WE). These values above as well as the average of SN and SP are used to measure the accuracy of a program.

RT-PCR and cDNA library screen to identify the yz1 gene

The five *ab initio* programs evaluated in this study predicted a gene, *yz1,* in the maize *a1-sh2* interval. To confirm this gene prediction and to determine the actual structure of this gene, RT-PCR experiments were conducted and maize cDNA libraries were screened.

SuperScript™ First-Strand Synthesis System for RT-PCR (Invitrogen, Carlsbad, CA) was used to obtain first-strand cDNA from total RNA. The sequences of oligonucleotides used as primers in the subsequent PCR are: YZ4b (5'- GAGATGATGTCCCTTGTG -3') and ZH2587 (5'- GCCTGGTTAGCGAAGTTG -3'). RT-PCR amplification using these two primers revealed a 681-bp fragment in maize RNA isolated from different organs, including husk, tassel, silk, adult leaf, ear and seedling (data not shown). The sequence of this RT-PCR fragment is identical to the predicted *yz1* exons and the predicted introns were missing from the RT-PCR product.

This RT-PCR product was used as a probe to screen maize cDNA libraries. A cross-hybridizing clone with a 2.1-kb insert was identified from a library prepared from seedlings of the inbred CI31A. Sequence analysis of this clone demonstrated that it is chimeric, with only 1.4-kb derived from the *yz1*. Sequence analysis of this cDNA clone, as well as 3'- and 5'- Rapid Amplification of cDNA Ends (RACE) (Invitrogen, Carlsbad, CA) experiments, suggested that this 1.4-kb sequence is full-length or nearly full-length.

Results

Evaluation of gene prediction programs for maize gene discovery

Five *ab initio* gene prediction programs (Table 1) were evaluated for their ability to predict maize genes from genomic sequences. The purpose of this evaluation was to help biologists select a strategy for the *ab initio* discovery of maize genes from partial genomic sequences using currently available tools. Hence, this evaluation did not seek to evaluate the algorithms per se upon which the gene prediction tools are based. Of the five evaluated gene prediction programs, Grail predicts splice sites and exons but not gene models; the remaining four

programs predict gene models as well as exons and splice sites. Most of the programs had been previously trained using monocots (e.g., maize and/or rice), but Grail was trained using *Arabidopsis*. Evaluations were conducted using a data set (data set 1) consisting of eight maize genomic gene sequences (Table 2) that could not have been included in the data sets used to train any of the five prediction programs because these sequences were released from GenBank only after the evaluation of the gene prediction programs had been completed.

Five of the gene sequences are full-length; one (*rf2d*) is partial. The GC contents of the genes (from start to stop codons) in data set 1 range from 39.9% to 61.5% with an average of 48.8%. The lengths of these gene sequences range from 2,899 to 13,621 bp (Table 2). There are 65 exons in this data set with one to 25 exons per gene. Exon lengths range from 62 to 2,004 bp with an average of 208 bp. The average intron length is 327 bp with a minimum of 70 bp and a maximum of 1,705 bp. The total numbers of donor and acceptor sites are 57 and 58, respectively. To determine if the genes in data set 1 are representative of maize genes as a whole, we compared several of their features to those of a set of 74 structure-known maize genes downloaded from GenBank (Methods). The Kolmogorov-Smirnov test (Kolmogorov, 1933; Smirnov, 1939) revealed no significant differences (P value > 0.2) in the lengths and GC contents of internal exons (i.e., those that begin with an acceptor site and end with a donor site) and introns (Figure 1). Therefore, data set 1 is at least reasonably representative of the maize genes that have been deposited in Genbank to date.

The performance of the five gene prediction programs was evaluated at three levels: splice site, nucleotide and exon. At the splice site level, the accuracy of a program is measured by the average value of SN and SP since neither SN nor SP alone is sufficient to

indicate the ability of a program to predict genes (Methods). FGENESH had the highest values of (SN + SP)/2 (Table 3). These values are 0.91 for donor sites and 0.92 for acceptor sites. GeneMark.hmm was the second most accurate program with values of (SN + SP)/2 = 0.84 for donor sites and 0.78 for acceptor sites. Both the SN and SP of FGENESH's predictions are high. In contrast, GeneMark.hmm exhibits a higher value of SP than SN. Such a differential is also present in the predictions from GENSCAN and GlimmerR which have SPs close to those of FGENESH and GeneMark.hmm but that have much lower SNs. GeneMark.hmm, GENSCAN and GlimmerR predict donor sites better than acceptor sites.

Accuracy at the nucleotide level (Table 4) is measured by the value of the correlation coefficient (CC, Methods). The programs, from the most accurate to the least measured by the CC, are FGENESH (0.93), GeneMark.hmm (0.89), GENSCAN (0.82), GlimmerR (0.71) and Grail (0.43). FGENESH has the highest SN (0.97) and GENSCAN has the highest SP (0.95). The values of SN and SP for FGENESH and GeneMark.hmm are both high (over 0.90). Although the SP values of GENSCAN and GlimmerR are also high (0.95 and 0.91, respectively), their SN values are less favorable (0.81 and 0.70, respectively).

At the exon level, the programs with values of (SN + SP)/2 from the highest to lowest are: FGENESH (0.87), GeneMark.hmm (0.75), GENSCAN (0.68), GlimmerR (0.57) and Grail (0.31). FGENESH has both the highest SN and SP (0.86 and 0.88, respectively). The SPs of GeneMark.hmm and GENSCAN (0.80 and 0.81, respectively) compare favorably with those of FGENESH but their SNs compare less favorably (0.69 and 0.54, respectively). GlimmerR also exhibits better SP than SN. Consistent with its highest SN and SP among the five evaluated programs, FGENESH has the lowest percentage of missing (ME% =4.6) and wrong (WE%=3.1) exons. Although the values of WE% in GeneMark.hmm, GENSCAN

and GlimmerR predictions are not high (5.4, 7.0 and 7.7, respectively), the values of ME% are 19, 39 and 23, respectively. Grail exhibited the lowest SN and SP and had correspondently high percentages of both MEs (ME%=17) and WEs (WE%=31). Predicted exons can have only one correct boundary (PE, partial exon) or can overlap the true exon but lack two correct boundaries (OE, overlapped exon). Of the exons predicted by FGENESH 9.4% and 0% were PEs and OEs, respectively. These are the lowest values of all evaluated programs. Predictions from GeneMark.hmm and GENSCAN also contain no OEs but 14% and 12% PEs, respectively. GlimmerR and Grail predicted both PEs and OEs, but the values of PE% are much higher than that of OE%.

Gene discovery in the a1-sh2 interval

To test the ability of the five evaluated programs to discover new maize genes, each was used to predict the structures of genes in the 15,783-bp (GenBank accession no. AF434192) and 6,506-bp fragments (GenBank accession no. AF434193) of the 140-kb maize *a1-sh2* interval (Yao *et al.*, 2002) (Figure 2). Because the *a1-sh2* sequences were not released to the public until after the completion of this evaluation, these sequences also could not have been included in the training sets of any of the prediction programs.

The Bennetzen lab (Chen and Bennetzen, 1996; Chen *et al.*, 1998) sequenced the *a1-sh2* intervals of rice and sorghum (GenBank accession no. U70541 and AF010283, respectively) and predicted a genic sequence between the *a1* and *sh2* loci, which they termed "*Gene X*". The identification of its maize homologue has been described by Yao *et al.* (2002). Comparison of the genomic and cDNA sequences revealed that the maize *x1* gene contains seven exons (Figure 2). Comparisons of the sequences of both the rice and maize

full-length *x1* cDNAs to the predicted rice or sorghum "*Gene X*" showed that only the 5' portion of the predicted "*Gene X*" corresponds to the actual rice and maize *x1* genes.

Comparisons of *a1-sh2* derived sequences from rice, sorghum and maize revealed a conserved region other than the *a1* and *x1* (Figure 2). This conserved region is located at the distal end of the 15,783-bp portion of the maize *a1-sh2* interval (GenBank accession no. AF434192) and overlaps with the 3'-portion of the predicted "*Gene X*" in rice and sorghum. Since the actual rice and maize *x1* genes do not contain this conserved region and part of this region is single-copy in the maize genome (Yao *et al.*, 2002), we hypothesized that there is another gene in the *a1-sh2* interval. To test this hypothesis, the five gene prediction programs were used to conduct predictions in a 5.4-kb segment from the distal end of the 15,783-bp sequence within the *a1-sh2* interval. All programs predicted a gene in this 5.4-kb sequence (Figure 2), although the predicted gene structures (or splice sites) vary. To confirm the validity of these predictions, primers (YZ4b and YZ2587) were designed in the putative exonic regions that were predicted by most programs and that exhibit a high degree of sequence similarity among rice, sorghum and maize (Figure 2). RT-PCR amplification using these primers and comparison of the sequence of the amplified fragment to the 5.4-kb genomic sequence revealed that as predicted by the *ab initio* programs, an additional expressed gene (termed as *yz1*) is present in the *a1-sh2* interval (Methods).

A 1.4-kb yz1 cDNA was isolated (Methods) that is nearly full-length, probably lacking only the five codons at its 5' end where a putative initiation Met resides. This putative initiation Met was predicted based on the fact that it and the following four amino acids are conserved among rice, sorghum and maize. Comparison of the 1.4-kb cDNA sequence of *yz1* to the 5.4-kb genomic sequence showed that the genomic sequence of *yz1* is

approximately 2.7 kb and consists of seven exons. In a more recently submitted *a1-sh2* sequence from the rice cultivar japonica (GenBank accession no. AF101045), the original "*Gene X*" has been annotated as *x1* and *x2* which are homologs of the maize *x1* and *yz1* genes, respectively.

Comparisons of predicted and actual yz1 and x1 splice sites and gene structures

Because a complete gene model predicted by FGENESH, GeneMark.hmm, GENSCAN and GlimmerR begins with the start codon and ends with the stop codon, the 5'- and 3'-UTRs of *yz1* and *x1* were not considered in the following comparisons.

FGENESH gave the second best prediction for the *yz1* gene and the best prediction for the *x1* gene at the splice site, nucleotide and exon levels (data not shown). Even so, none of the gene models predicted by FGENESH, GeneMark.hmm, GENSCAN and GlimmerR for *yz1* and *x1* is completely correct (Figure 2). The start and stop codons of *yz1* are located in the first and the last (seventh) exons, respectively. Although FGENESH, GeneMark.hmm and GENSCAN correctly predicted the positions of the stop codons, each of these programs missed the start codon by predicting the first exon as internal rather than initial (i.e., one that starts with the initiation codon and ends with a donor site). The start and stop codons of *x1* are located in the second and last (seventh) exons, respectively. None of the four programs correctly predicted the position of the start codon. Whereas FGENESH, GENSCAN and GlimmerR correctly predicted the location of the stop codon, GeneMark.hmm's prediction is incorrect.

A particular problem with predicting maize genes using the version of GlimmerR that was trained for rice is that it splits maize genes. As shown in Figure 2, GlimmerR predicted

multiple genes using the *yz1* and *x1* gene sequences. It predicted three genes in both the *yz1*- and the *x1*-containing sequences: two of these predicted genes each consist of a single exon (which starts with an initiation codon and ends with a stop codon); while in each case the other predicted gene contains multiple exons. GlimmerR was used to predict genes in genomic sequences from the rice *a1-sh2* interval that correspond to the *x1* and *yz1* genes. Althought the sensitivity of GlimmerR in predicting *x1* and *yz1* was not improved by using rice sequences (data not shown), no split genes were predicted (data not shown). Hence, the observed gene splitting could be a consequence of using a version of GlimmerR that had been trained on rice to predict maize genes.

Gene predictions in MAGIs

Gene prediction programs are of particular importance in predicting genes in large genome projects. We therefore extended our evaluations to the maize GSSs being generated as part of the NSF Plant Genome project 0221536 and assembled into MAGIs at Iowa State University (http://plantgenomics.iastate.edu/maize). Data set 2 consists of 1,353 MAGI contigs that aligned well with B73 3' ESTs sequenced by us and that contain at least one pair of reliable donor and acceptor sites flanking an intact intron (Figure 3 and Supplementary Materials). There are 1,928 pairs of reliable canonical splice sites and 18 pairs of reliable non-canonical splice sites (16 GC-AG pairs, 2 AT-AC pairs) in this data set that correspond to 592 reliable exons and 1,946 reliable introns. Detailed statistical characteristics of data set 2 are provided in Figure 1.

Data set 2 was analyzed with only FGENESH, GeneMark.hmm and GENSCAN because these programs were trained using maize sequences and proved more reliable in the

analyses of data set 1 than other two programs. Predictions of data set 2 from each of the three tested programs were parsed for subsequent analysis. Only regions flanked by two reliable splice sites and without the presence of non-reliable internal splice sites were considered in the evaluation (Figure 3). As was done for data set 1, the accuracy of the predictions by each of the three programs was evaluated at the splice site, nucleotide and exon levels (Methods). As shown in Table 5, the overall accuracy of each the program was somewhat reduced as compared to that obtained using data set 1 (Tables 3 and 4) due to the decreased specificities of the predictions at the nucleotide and exon levels and decreased sensitivities at all three levels. At the splice site level, the specificities of the predictions by the three programs remained high and indeed increased somewhat. Overall, FGENESH performed better than the other two programs because of its much higher sensitivity, even though its specificity is slightly lower than that GeneMark.hmm (Table 5).

At the nucleotide level, FGENESH's values of SN and SP are 0.86 and 0.84, respectively with a CC of 0.83. Consistently, when considering only the MAGIs that have FGENESH predictions at the regions evaluated in our analyses (Figure 3), SN and SP are well correlated at the nucleotide level; this reflects the fact that the majority of these MAGIs have SN and SP values equal to 1 (Data not shown). Therefore, if a MAGI is predicted by FGENESH to contain a gene, that prediction is likely to be correct.

When running FGENESH to predict genes in data set 2, the "–GC" parameter was used. This allows FGENESH to predict non-canonical GC donor sites. FGENESH correctly predicted 13/16 (81%) of the non-canonical GC donor sites in data set 2. In contrast, none of these GC donor sites were correctly predicted by GeneMark.hmm and GENSCAN. Neither

of the two pairs of non-canonical AT-AC splice sites in the data set 2 was identified by any of the three programs.

FGENESH, GeneMark.hmm and GENSCAN correctly predicted the gene models in 773, 625, and 371 MAGIs, respectively, out of the 1,353 MAGIs in data set 2 (Figure 4). FGENESH, GeneMark.hmm and GENSCAN uniquely and correctly predicted 214, 94 and 21 MAGIs, respectively. FGENESH, GeneMark.hmm and GENSCAN failed to predict the evaluated regions as genic in 249, 235, and 540 MAGIs, respectively. FGENESH, GeneMark.hmm and GENSCAN uniquely missed the evaluated genic segments completely in 50, 31 and 275 MAGIs, respectively.

If the predictions from all three programs are considered together, the number of correctly predicted MAGIs increases to 911and the numbers of MAGIs that were completely missed drops to 112. These results suggest that combining the prediction results from different programs can increase the accuracy of predictions.

Comparisons of predictions of internal versus initial/terminal exons

In vertebrate and *Drosophila* genomic sequences FGENESH and GENSCAN predict internal exons better than they predict initial and terminal exons (i.e., those that begin with an acceptor site and end with a stop codon) (Burge and Karlin, 1998; Salamov and Solovyev, 2000). This reflects the poorer abilities of these programs to detect the correct start and stop codons than their abilities to correctly identify splice sites. The abilities of the four programs evaluated in this study, FGENESH, GeneMark.hmm, GENSCAN and GlimmerR, to predict initial/terminal exons versus internal exons were compared at the exon level using the eight

genes in data set 1 and the *yz1* and *x1* genes from the *a1-sh2* interval. Grail was not included in this analysis because it does not predict exons as initial, internal and terminal.

As expected based on experience from other genomes, FGENESH, GeneMark.hmm and GlimmerR predicted the internal exons better than the initial and terminal exons (Table 6). Surprisingly, GENSCAN predicted initial and terminal exons better than it did internal exons due to its higher SN of the initial and terminal exons in data set 1.

Selecting reliable exon prediction

The accuracy of gene prediction at the exon level, as well as the prediction of gene models is not as high as the accuracy at the nucleotide level. This is because more than 12% of the predicted exons are PEs, OEs and WEs (Tables 4 and 5). In applications such as primer design for RT-PCR experiments or the design of oligos for microarrays, it is highly desirable to be able to exclude WE predictions. Of the two most reliable programs in this evaluation, FGENESH and GeneMark.hmm, only FGENESH reports a confidence score for its exonic predictions. This score is an aggregate of log-odds scores that the base pairs are members of an exon. Unfortunately, we have been unable to locate in FGENESH documentation or its references the method used to estimate the context-sensitive probability that a base is a member of an exon or the method of aggregating these scores into an overall exon score.

To determine if an FGENESH exon score correlates with the quality of prediction, the distributions of exon's scores were compared among the TEs, PEs+OEs and WEs (Figure 5) from predictions using data set 2. About 49% of the WEs have negative scores. In contrast, only 2.2% of the TEs and 6.6% of the PEs and OEs have negative scores. Considering only

predicted exons with non-negative scores, 68% are TEs, 27% are PEs+OEs, and only 4.5% are WEs. These results demonstrate that removing exons with negative FGENESH scores eliminates almost half of the WEs, while retaining the majority of the TEs and PEs+OEs. Indeed, in data set 2 all WEs have scores of less than 10 (Figure 5). Hence, if only those predicted exons with scores greater than 10 were used WEs could be totally eliminated.

To determine the effect of removing exons with negative scores on the evaluation of FGENESH, the parameters for the evaluation of FGENESH's predictions at the splice site, nucleotide and exon levels were recalculated for data set 2 (Table 5, FGENESH (Score $\geq$0)). In this analysis, predicted exons with negative scores were treated as they had not been predicted. This modified analysis resulted in higher SPs at all three levels as compared to the SPs obtained in the original analysis of FGENESH. In contrast, SNs decreased. The values of (SN+SP)/2 for splice site and exon levels, CC, and ME% and WE% were altered only slightly.

Discussion

FGENESH performed better than other evaluated programs for maize gene discovery

The goal of this study was to identify a strategy based on existing gene prediction *ab initio* gene prediction tools that biologists can use to discover maize genes in genomic sequences. Accordingly the performances of five *ab initio* gene prediction programs were evaluated using data set 1, which consists of eight maize genes that could not have been used to train these programs. These eight genes are structurally similar to a larger set of 74 structure-known genes downloaded from GenBank (Figure 1). Hence, evaluation of predictions

performed on the eight genes in data set 1 are likely to be informative of the ability of these programs to predict other maize genes.

In these evaluations FGENESH performed the best at all three levels of evaluation. FGENESH has also been demonstrated to be more accurate than other tested programs for the discovery of rice and mammalian genes (Yu *et al.*, 2002; Solovyev, 2001). GeneMark.hmm is the most accurate program for *Arabidopsis* gene discovery when evaluated using the AraSet that contains contigs of validated genes (Pavy *et al.*, 1999). In our evaluation, GeneMark.hmm was the second most accurate program. GENSCAN, which is very good at predicting mammalian genes (Rogic *et al.*, 2001), fared less well in our analysis of maize genes. This may be due to the fact that GENSCAN was trained on a smaller data set than FGENESH and GeneMark.hmm (Supplementary Materials). One reason for the poor performance of GlimmerR and Grail in predicting the maize genes may be because they had been trained for other plants (including rice and *Arabidopsis*, Table 1). Since gene features differ among organisms, parameters of these programs may not be optimized for maize gene discovery. Additional evaluations of FGENESH, GeneMark.hmm and GENSCAN using data set 2, which consists of 1,353 genic MAGIs, also demonstrated that FGENESH is most accurate at predicting maize genes. It is, however, important to emphasize that this study was not designed to evaluate the algorithms used by these programs. On the other hand, this study does reveal which existing programs will provide maize biologists with the best gene predictions.

Predictions of small exons may be less accurate

The accuracies of FGENESH, GeneMark.hmm and GENSCAN predictions in data set 2 are not as high as those in data set 1. This may be due the increased fraction of small exons (e.g., $\leq$ 100 bp) in data set 2 as compared to data set 1 (Figure 1). This enrichment for small exons in data set 2 is probably a consequence of our stringent EST-guided strategy to select reliable genic regions in MAGIs for analyses (Figure 3, Methods and Supplementary Materials). It has been demonstrated that in rice genes FGENESH is not as successful at predicting small exons (less than 200 bp) as large exons (Yu *et al.*, 2002). The finding that in data set 2 the fractions of MEs in exons that are smaller than 50 bp is significantly higher than the corresponding fraction among larger (> 50 bp) exons (χ^2 test, p value = 0.002) is consistent with the hypothesis that this enrichment for small exons is at least partly responsible for the reduced accuracy of predictions in data set 2 as compared to those of data set 1.

Gene model prediction programs need improvement

As shown in Table 7, none of the four gene model prediction programs (FGENESH, GeneMark.hmm, GENSCAN and GlimmerR) precisely predicted the structures of more than half of the eight genes in data set 1 plus the two genes from the *a1-sh2* interval. The structures of four genes (*rf2e1*, *x1*, *rth1*, and *yz1*) were not predicted correctly by any of the four programs. These programs each appear to have difficulties predicting gene models that include non-canonical splice sites, start codons that are not in the first exon, or large numbers of small exons and/or large introns. For example, the inability of these programs to correctly predict non-canonical splice sites appears to be the reason they failed to predict correctly the gene model of *rf2e1*. The *rf2e1* gene contains two pairs of non-canonical splice sites,

GC/AG and CC/AA, in introns 2 and 3, respectively. Each of the four programs missed both of these two non-canonical donor and acceptor sites.

The start codon of the *x1* gene is in its second exon (Figure 2), which may interfere with the ability of prediction programs to identify it. The reason for the incorrect prediction of the start codon in the *yz1* gene is not clear. The *rth1* gene has 25 exons, each of which is less than the average length of maize exons (i.e., 200 bp, Table 2). Thirteen of the *rth1* exons are between 50 and 100 bp, eight are between 100 and 150 bp and the remaining four are between 150 and 200 bp in length. In contrast, eight of the *rth1* introns are larger than the average maize intron (i.e., 300 bp, Table 2) and four are over 900 bp. All of the four assayed programs missed some of *rth1*'s small exons and incorrectly predicted the presence of exons within *rth1*'s large introns. Three of the programs (GENSCAN, GeneMark.hmm and GlimmerR) even split the *rth1* gene, which may indicate a poor ability to predict large genes. As pointed out by Wang *et al.* (2003) because *ab initio* programs predict genes based on statistical analyses of all possible genic features (e.g., splice sites, start and stop codons), longer sequences have an increased probability on containing false genic features that exhibit statistical significance. In addition, stop codons are more likely to be associated with FP predictions in intron, which could split large genes (which usually contain large introns). Our study provides additional evidence that GlimmerR's predictions tend to incorrectly split maize genes. GlimmerR split five genes (*gl8a*, *rf2e1*, *rth1*, *yz1* and *x1*). Since GlimmerR was trained for rice, the current version may not suitable for the prediction of maize genes. We conclude that *ab initio* gene model prediction remains a field that would benefit from further research.

The ability of FGENESH to predict non-canonical splice sites

Non-canonical splice sites can make the accurate prediction of gene models difficult because until recently no program was trained to recognize non-canonical splice sites due to an insufficient number of non-canonical sites in the training sets. As more genomic sequences have become available, data sets of EST-supported canonical and non-canonical mammalian splice sites have been created and analyzed (Burset *et al.*, 2000; Burset *et al.*, 2001). In these mammalian splice site data sets, the canonical GT-AG pairs account for 98.7% of all splice site pairs; non-canonical GC-AG pairs and AT-AC pairs account for 0.56% and 0.05%, respectively, and all other non-canonical pairs account for 0.02%. The collection of GC-AG pairs in this mammalian data set was large enough for training and an updated version of FGENESH (for mammals) incorporates GC donor sites in its predictions.

Analysis of spliced alignments between clustered *Arabidopsis* EST and genomic sequences also showed that the canonical GT-AG pairs account for the majority of the splice sites in *Arabidopsis* (Zhu *et al.*, 2003). In that species the frequencies of the non-canonical GC-AG and AT-AC sites have been estimated to be about 1.0% and 0.06%, respectively. These may, however be over-estimates because ambiguous splice sites were included in this analysis (Zhu *et al.*, 2003). In our data set 2, 99.1% of all sites were canonical GT-AG pairs and non-canonical GC-AG and AT-AC pairs represent 0.822% and 0.103% of all pairs, respectively. This result indicates that the fractions of non-canonical GC-AG pairs in maize and *Arabidopsis* and AT-AC pairs in maize may be higher than in mammalian genomes.

By using the "–GC" parameter, FGENESH was able to identify 81% (13/16) of the non-canonical GC donor sites in data set 2. Since most donor sites in data set 2 are canonical and the sensitivity of them is 0.73, FGENESH's sensitivity for non-canonical GC sites is at least as good as it's sensitivity for canonical GT sites.

Recommendations for gene prediction

Of the evaluated *ab initio* programs FGENESH provided the highest degree of SP and SN, followed by GeneMark.hmm. Both of these programs provide high levels of SP with acceptable (but somewhat lower) levels of SN. Consequently, if a sequence is predicted to contain a gene, that prediction is likely to be correct, but some sequences that do contain genes will be missed. Using its "–GC" parameter, FGENESH is able to identify many non-canonical GC donor sites. Removing exons with negative FGENESH scores will eliminate most of the WEs, while retaining the majority of the TEs and PEs+OEs. Therefore, for RT-PCR experiments and microarray design projects it is better to avoid designing primers or oligos in predicted exons with negative scores. If the specificity of exon prediction is the priority, predicted exons with even higher scores (≥ 10) should be used. Although will result in the loss of correct exons, it will also eliminate essentially all wrong exons.

Combining gene prediction results from multiple *ab initio* programs improves gene model predictions (reviewed by Mathe *et al.*, 2002) because even a good program can make incorrect predictions for some genes and even a poor program can make correct predictions for some genes. For example, as shown in Table 7, FGENESH did not correctly predict the gene model of *pdc3*, but the other three programs did. Moreover, analysis of predictions of

genes in data set 2 suggests that by considering predictions from FGENESH, GeneMark.hmm and GENSCAN, it is possible to improve the accuracy of *ab initio* gene discovery (Figure 4). Integration of *ab initio* and sequence similarity based approaches is another way to improve the accuracy of gene prediction and is likely to be more widely used as the number of sequences genomes increases (reviewed by Mathé *et al.*, 2002). The Twinscan (Korf *et al.*, 2001) and Combiner (Allen *et al.*, 2004) programs improve the accuracy of gene predictions via theses two approaches. Development of similar programs or training available programs for maize sequences could also contribute to the efficient discovery of maize genes.

Acknowledgments

The *rth3*, *rf2e1*, *pdc2* and *pdc3* genes were cloned using cDNA clones identified via searches of Pioneer Hi-Bred's proprietary EST database. We thank Tim Fox, Wes Bruce and Carl Simmons (Pioneer Hi-Bred, Intl. Inc., Johnston, IA) for their assistance with these searches. We thank Heike Hofmann (Iowa State University) for assistance with the Kolmogorov-Smirnov tests and Volker Brendel (Iowa State University) for helpful discussions. This research was funded in part by competitive grants from the National Science Foundation Plant Genome Program (awards: DBI-9975868, DBI-0121417, and DBI-0321711) and the United States Department of Agriculture National Research Initiative Program (awards: 98101805, 0001478, 0201419, and 0300940). Support was also provided by Hatch Act and State of Iowa funds.

References

Allen, J.E., Pertea, M. and Salzberg, S.L. 2004 Computational gene prediction using multiple sources of evidence. Genome Res. 14: 142-148.

Bennetzen, J.L., Chandler, V.L. and Schnable, P.S. 2001. National Science Foundation-sponsored workshop report. Maize genome sequencing project. Plant Physiol. 127: 1572-1578.

Brendel, V., Xing, L. and Zhu, W. 2004 Gene structure prediction from consensus spliced alignment of multiple ESTs matching the same genomic locus. Bioinformatics 20: 1157-1169.

Burge, C. and Karlin, S. 1997. Prediction of complete gene structures in human genomic DNA. J. Mol. Biol. 268: 78-94.

Burge, C. and Karlin, S. 1998. Finding the genes in genomic DNA. Curr. Opin. Struct. Biol. 8: 346-354.

Burset, M. and Guigo, R. 1996. Evaluation of gene structure prediction programs. Genomics 34: 353-367.

Burset, M., Seledtsov, I.A. and Solovyev, V.V. 2000. Analysis of canonical and non-canonical splice sites in mammalian genomes. Nucleic Acids Res. 28: 4364-4375.

Burset, M., Seledtsov, I.A. and Solovyev, V.V. 2001. SpliceDB: database of canonical and non-canonical mammalian splice sites. Nucleic Acids Res. 29: 255-259.

Civardi, L., Xia, Y., Edwards, K.J., Schnable, P.S. and Nikolau, B.J. 1994. The relationship between genetic and physical distances in the cloned *a1-sh2* interval of the *Zea mays* L. genome. Proc. Natl. Acad. Sci. USA 91: 8268-8272.

Chen, M. and Bennetzen, J.L. 1996. Sequence composition and organization in the *Sh2/A1*-homologous region of rice. Plant Mol. Biol. 32: 999-1001.

Chen, M., SanMiguel, P. and Bennetzen, J.L. 1998. Sequence organization and conservation in *sh2/a1*-homologous regions of sorghum and rice. Genetics 148: 435-443.

Emrich, S.J., Aluru, S., Fu, Y., Wen, T.-J., Narayanan, M., Guo, L., Ashlock, D.A., Schnable, P.S. 2004. Astrategy for assembling the maize (*Zea mays* L.) genome. Bioinformatics 20: 140-147.

Goff, S.A. *et al.* 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). Science 296: 92-100.

Hebsgaard, S.M., Korning, P.G., Tolstrup, N., Engelbrecht, J., Rouze, P. and Brunak, S. 1996. Splice site prediction in Arabidopsis thaliana pre-mRNA by combining local and global sequence information. Nucleic Acids Res. 24: 3439-3452.

Korf, I., P. Flicek, D. D. and Brent, M.R. 2001. Integrating genomic homology into gene structure prediction. Bioinformatics 17: 140-148.

Kolmogorov, A.N. 1933. Sulla determinazione empirica di una legge di distribuzione. Giornale dell' Istituto Italiano degli Attuari 4: 83-91.

Lukashin, A.V. and Borodovsky, M. 1998. GeneMark.hmm: new solutions for gene finding. Nucleic Acids Res. 26: 1107-1115.

Mathé, C., Sagot, M.F., Schiex, T. and Rouze, P. 2002. Current methods of gene prediction, their strengths and weaknesses. Nucleic Acids Res. 30: 4103-4117.

Moore, G. 2000. Cereal chromosome structure, evolution, and pairing. Annu. Rev. Plant Physiol. Plant Mol. Biol. 51: 195-222.

Pavy, N., Rombauts, S., Dehais, P., Mathe, C., Ramana, D.V., Leroy, P. and Rouze, P. 1999. Evaluation of gene prediction software using a genomic data set: application to Arabidopsis thaliana sequences. Bioinformatics 15: 887-899.

Palmer, L.E., Rabinowicz, P.D., O'Shaughnessy, A.L., Balija, V.S., Nascimento, L.U., Dike, S., de la Bastide, M,, Martienssen, R.A. and McCombie, W.R. 2003. Maize genome sequencing by methylation filtration. Science 302: 2115-2117.

Pertea, M., Lin, X. and Salzberg, S.L. 2001. GeneSplicer : a new computational method for splice site prediction. Nucleic Acids Res. 29: 1185-1190.

Pertea, M. and Salzberg, S.L. 2002. Computational gene finding in plants. Plant Mol. Biol. 48: 39-48.

Peterson, D.G., Wessler, S.R. and Paterson, A.H. 2002. Efficient capture of unique sequences from eukaryotic genomes. Trends Genet. 18: 547-550.

Rabinowicz, P.D., Schutz, K., Dedhia, N., Yordan, C., Parnell, L.D., Stein, L., McCombie, W.R. and Martienssen, R.A. 1999. Differential methylation of genes and retrotransposons facilitates shotgun sequencing of the maize genome. Nature Genetics 23: 305–308.

Rogic, S., Mackworth, A.K. and Ouellette, F.B.F. 2001. Evaluation of Gene-Finding Programs on Mammalian Sequences. Genome Res. 11: 817-832.

Salamov, A.A. and Solovyev, V.V. 2000. *Ab initio* gene finding in Drosophila genomic DNA. Genome Res. 10: 516-522.

Salzberg, S.L., Pertea, M., Delcher, A.L., Gardner, M.J. and Tettelin, H. 1999. Interpolated Markov models for eukaryotic gene finding. Genomics 59: 24-31.

SanMiguel, P., Tikhonov, A., Jin, Y.K., Motchoulskaia, N., Zakharov, D., Melake-Berhan, A., Springer, P.S., Edwards, K.J., Lee, M., Avramova, Z. and Bennetzen, J.L. 1996. Nested retrotransposons in the intergenic regions of the maize genome. Science 274: 765-768.

Smirnov, N.V. 1939. Estimate of deviation between empirical distribution functions in two independent samples. Bulletin Moscow University 2: 3-16.

Solovyev, V. 2001. Statistical approaches in eukaryotic gene prediction. In: D.J. Balding, M. Bishop and C. Cannings (Ed.), Handbook of Statistical Genetics, John Wiley & Sons, Ltd, New York, pp. 83-127.

Stormo, G.D. 2000. Gene-finding approaches for eukaryotes. Genome Res. 10: 394-397.

The Arabidopsis Genome Initiative 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408: 796 – 815.

Tolstrup, N., Rouze, P. and Brunak, S. 1997. A branch point consensus from Arabidopsis found by non-circular analysis allows for better prediction of acceptor sites. Nucleic Acids Res. 25: 3159-3163.

Usuka, J. and Brendel, V. 2000. Gene structure prediction by spliced alignment of genomic DNA with protein sequences: increased accuracy by differential splice site scoring. J. Mol. Biol. 297: 1075–1085.

Usuka, J., Zhu, W. and Brendel, V. 2000. Optimal spliced alignment of homologous cDNA to a genomic DNA template. Bioinformatics 16: 203–211.

Wang, J., Li, S., Zhang, Y., Zheng, H., Xu, Z., Ye, J., Yu, J. and Wong, G.K. 2003. Vertebrate gene predictions and the problem of large genes. Nat. Rev. Genet. 4: 741-749.

Whitelaw, C.A., Barbazuk, W.B., Pertea, G., Chan, A.P., Cheung, F., Lee, Y., Zheng, L., van Heeringen, S., Karamycheva, S., Bennetzen, J.L., SanMiguel, P., Lakey, N., Bedell, J., Yuan, Y., Budiman, M.A., Resnick, A., Van Aken, S., Utterback, T., Riedmuller, S., Williams, M., Feldblyum, T., Schubert, K., Beachy, R., Fraser, C.M. and Quackenbush, J. 2003. Enrichment of gene-coding sequences in maize by genome filtration. Science 302: 2118-2120.

Xu, Y. and Uberbacher, E.C. 1997. Automated gene identification in large-scale genomic sequences. J. Comput. Biol. 4: 325-38.

Yao, H., Zhou, Q., Li, J., Smith, H., Yandeau, M., Nikolau, B.J. and Schnable, P.S. 2002. Molecular characterization of meiotic recombination across the 140-kb multigenic *a1-sh2* interval of maize. Proc. Natl. Acad. Sci. USA. 99: 6157-6162.

Yu, J. *et al.* 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). Science 296: 79-92.

Yuan, Q., Quackenbush, J., Sultana, R., Pertea, M., Salzberg, S.L. and Buell, C.R. 2001. Rice bioinformatics. Analysis of rice sequence data and leveraging the data to other plant species. Plant Physiol. 125: 1166-1174.

Yuan, Y., SanMiguel, P.J. and Bennetzen, J.L. 2003. High-C_0t sequence analysis of the maize genome. Plant J. 34: 249-255.

Zhu, W., Schlueter, S.D. and Brendel, V. 2003. Refined annotation of the Arabidopsis genome by complete expressed sequence tag mapping. Plant Physiol. 132: 469-484.

Figure legends

Figure 1. The GC contents and lengths of exons and introns in data sets 1, 2 and structure-known genes. Only internal exons were analyzed in data set 1 and the structure-known genes. The percentages of exons (panel A) and introns (panel B) with the indicated GC contents (bin sizes = 5 percentage points) in each of three data sets are indicated. The percentages of exons (panel C) and introns (panel D) with the indicated lengths (bins sizes = 50 bp) in each of three data sets are also indicated. Data set 1, structure-known genes, and data set 2, are indicated by horizontal stripes, dark gray fill and diagonal stripes respectively.

Figure 2. Gene discovery in the *a1-sh2* interval of maize. The gene structures of the *yz1* and *x1* genes predicted by the five indicated programs and their actual structures as verified via RT-PCR and sequencing of cDNA clones are shown. The positions of initiation Mets (M) in the actual genes and predicted gene models are shown. The positions of stop codon are designated by *. Gray regions are conserved among rice, sorghum and maize. In the gene

models predicted by GlimmerR, exons filled with different patterns belong to different predicted genes. Primers used in RT-PCR are shown as horizontal arrows.

Figure 3. Criteria used to select qualified alignments for data set 2. ESTs that contained polyA tails of at least 8 A's were aligned to MAGI contigs. Alignments between a genomic sequence and an EST can be either terminal or internal. To qualify, terminal alignments (TA) between a MAGI contig and EST must be >= 50 bp with >= 98% nucleotide identity; internal alignments (IA) must be flanked by two qualified terminal alignments and exhibit >= 98% nucleotide identity. In addition, the 10 bp alignments at each splice junctions (shaded regions) must exhibit 100% nucleotide identity. It is possible that the end points of the alignment may not be the real boundaries of exons due to the incompleteness of the MAGI contig (e.g., 3' end) or the EST (e.g., the 5' end). These end points were therefore masked and not used to evaluate FGENESH, GeneMark.hmm and GENSCAN. Although not shown in this figure there are MAGI contigs that contain only two qualified terminal alignments and MAGI contigs that contain more than one qualified internal alignments. M's: masked end points of an alignment between a MAGI contig and EST; D's, donor sites; A's, acceptor sites.

Figure 4. Numbers of MAGIs correctly predicted, missed or predicted completely incorrectly by FGENESH, GeneMark.hmm and GENSCAN. In these comparisons, the entire gene model in the evaluated region of each MAGI (Figure 3) was considered. Predictions that exactly matched the actual gene model were classified as correctly predicted. Predictions that failed to identify the evaluated region as genic were classified as missed. Genic predictions that failed to correctly identify any features of the actual gene model were classified as completely incorrect. F, FGENESH; GM, GeneMark.hmm; GS, GENSCAN.

Figure 5. The distributions of exon scores among the TE (true exon), PE+OE (partial and overlapped exon) and WE (wrong exon) predicted by FGENESH using data set 2.

Table 1. Evaluated gene prediction programs.

Programs	Web Sites	Trained Organisms	Type of Prediction			Algorithm Models
			Splice Site	Exon	Gene Model	
FGENESH	http://www.softberry.com/berry.phtml?topic=gfind&prg=FGENESH	Monocots	Yes	Yes	Yes	GHMM[a]
GeneMark.hmm	http://opal.biology.gatech.edu/GeneMark/eukhmm.cgi?org=H.sapiens	Maize	Yes	Yes	Yes	GHMM
GENSCAN	http://genes.mit.edu/GENSCAN.html	Maize	Yes	Yes	Yes	GHMM
GlimmerR	http://www.tigr.org/tdb/glimmerm/glmr_form.html	Rice	Yes	Yes	Yes	IMM[b]
Grail	http://compbio.ornl.gov/Grail-1.3/	Arabidopsis	Yes	Yes	No	Neural Networks

[a]GHMM, Generalized Hidden Markov Model. [b]IMM, Interpolated Markov Model.

Table 2. Members and characteristics of data set 1.

Genes	GenBank Accession Numbers	(G+C)% of Gene[a]	Input Sequence Length (bp)	#D[b]	#A[c]	#Exons	Exon Length (bp)			Intron Length (bp)		
							Min	Max	Average	Min	Max	Average
gl8a	AF302098	50.0	3288	2	2	3	70	653	327	583	829	706
pdc2	AF370004	51.2	3974	5	5	6	118	651	297	82	691	259
pdc3	AF370006	54.1	3477	5	5	6	118	651	304	77	391	152
rf2c	AF348412	56.1	4527	6	6	7	62	648	216	70	1604	354
rf2d[d]	AF348414	54.9	2940	6	7	7	62	474	200	72	123	96
rf2e1	AY374447	54.3	4673	9	9	10	69	237	134	75	1080	271
rth1	AY265854	39.9	13621	24	24	25	65	174	107	80	1705	419
rth3	AY265855	61.5	2899	0	0	1	2004	2004	2004	NA[e]	NA	NA
Overall		48.8	39399	57	58	65	62	2004	208	70	1705	327

[a]Beginning with the start codon and ends with the stop codon.
[b]#D, number of donor sites. [c] #A, number of acceptor sites.
[d]The *rf2d* gene sequence is partial in the 5' end. The first intron is partial and was not included for analysis of intron length here although the A site of this intron is included for counting the #A.
[e]NA, Not applicable.

Table 3. The accuracy of gene predictions in data set 1 at the splice site level.

	Donor Sites			Acceptor Sites		
Programs	**SN**	**SP**	**(SN+SP)/2**	**SN**	**SP**	**(SN+SP)/2**
FGENESH	0.91	0.91	0.91	0.91	0.93	0.92
GeneMark.hmm	0.77	0.92	0.84	0.71	0.85	0.78
GENSCAN	0.56	0.91	0.74	0.53	0.86	0.70
GlimmerR	0.61	0.95	0.78	0.59	0.92	0.75
Grail	0.49	0.39	0.44	0.66	0.51	0.58

Table 4. The accuracy of gene predictions on data set 1 at the nucleotide and exon levels.

	Nucleotide Level			Exon Level						
Programs	**SN**	**SP**	**CC**	**SN**	**SP**	**(SN+SP)/2**	**PE%**	**OE%**	**ME%**	**WE%**
FGENESH	0.97	0.94	0.93	0.86	0.88	0.87	9.4	0	4.6	3.1
GeneMark.hmm	0.92	0.93	0.89	0.69	0.80	0.75	14	0	19	5.4
GENSCAN	0.81	0.95	0.82	0.54	0.81	0.68	12	0	39	7.0
GlimmerR	0.70	0.91	0.71	0.51	0.64	0.57	23	5.8	23	7.7
Grail	0.55	0.67	0.43	0.34	0.28	0.31	33	7.7	17	31

Table 5. The accuracy of gene predictions in data set 2.

	Donor Sites			Acceptor Sites			Nucleotide Level			Exon Level					
	SN	SP	(SN+SP) 2	SN	SP	(SN+SP) 2	SN	SP	CC	SN	SP	(SN+SP) 2	PE+ OE	WE	ME
GENSCAN	0.41	0.95	0.68	0.39	0.92	0.66	0.46	0.86	0.60	0.33	0.57	0.45	36%	7.6%	57%
GeneMark. hmm	0.69	0.94	0.82	0.63	0.93	0.78	0.77	0.88	0.80	0.66	0.67	0.66	28%	5.9%	20%
FGENESH	0.73	0.95	0.84	0.71	0.92	0.82	0.86	0.84	0.83	0.74	0.65	0.69	27%	8.2%	14%
FGENESH (score ≥ 0)[a]	0.71	0.95	0.83	0.67	0.94	0.81	0.83	0.86	0.82	0.72	0.68	0.70	27%	4.4%	17%

[a]FGENESH (score ≥ 0) is a modified evaluation of the FGENESH's prediction, in which predicted exons with negative scores were treated as if there were no predictions and which were therefore not included in the evaluation.

Table 6. Comparisons of accuracy at the exon level between the prediction of initial/terminal exons and internal exons.

	Initial and Terminal Exon							Internal Exon						
Programs	**SN**	**SP**	**$\frac{(SN+SP)}{2}$**	**PE%**	**OE%**	**ME%**	**WE%**	**SN**	**SP**	**$\frac{(SN+SP)}{2}$**	**PE%**	**OE%**	**ME%**	**WE%**
FGENESH	0.77	0.72	0.74	22	0	0	5.6	0.87	0.85	0.86	8.2	0	5.0	6.6
GeneMark.hmm	0.71	0.60	0.65	20	5.0	0	15	0.67	0.80	0.73	14	0	22	6.0
GENSCAN	0.71	0.63	0.67	16	0	12	21	0.42	0.83	0.63	13	0	52	3.3
GlimmerR	0.47	0.44	0.46	17	11	29	28	0.48	0.66	0.57	25	6.8	25	2.3

Table 7. Comparisons of gene model predictions.

Programs	***gl8a***	***pdc2***	***pdc3***	***rf2c***	***rf2d***	***rf2e1***	***rth1***	***rth3***	***x1***	***yz1***	**Number (%) of Correct Models**
FGENESH	Y[a]	Y	N	Y	Y	N	N	Y	N	N	5 (50%)
GeneMark.hmm	Y	N	Y	Y	N	N	N	Y	N	N	4 (40%)
GENSCAN	Y	Y	Y	Y	Y	N	N	N	N	N	5 (50%)
GlimmerR	N[b]	N	Y	N	N	N	N	N	N	N	1 (9%)
Number of Programs that Predicted Correct Models	3	2	3	3	2	0	0	2	0	0	

[a]Y, Prediction of the gene model is correct. [b]N, prediction of gene model is incorrect.

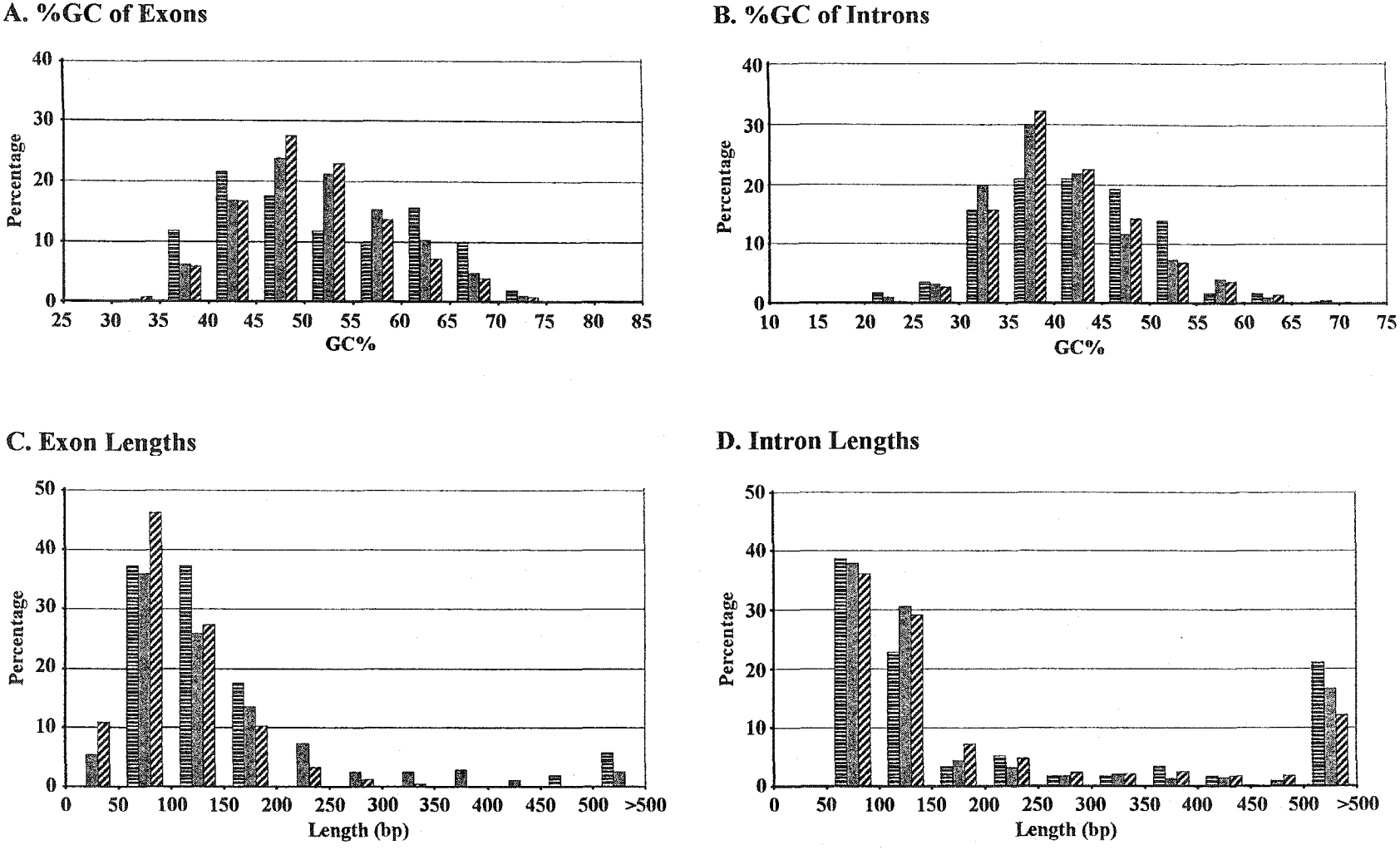


Figure 1. Yao *et al.*

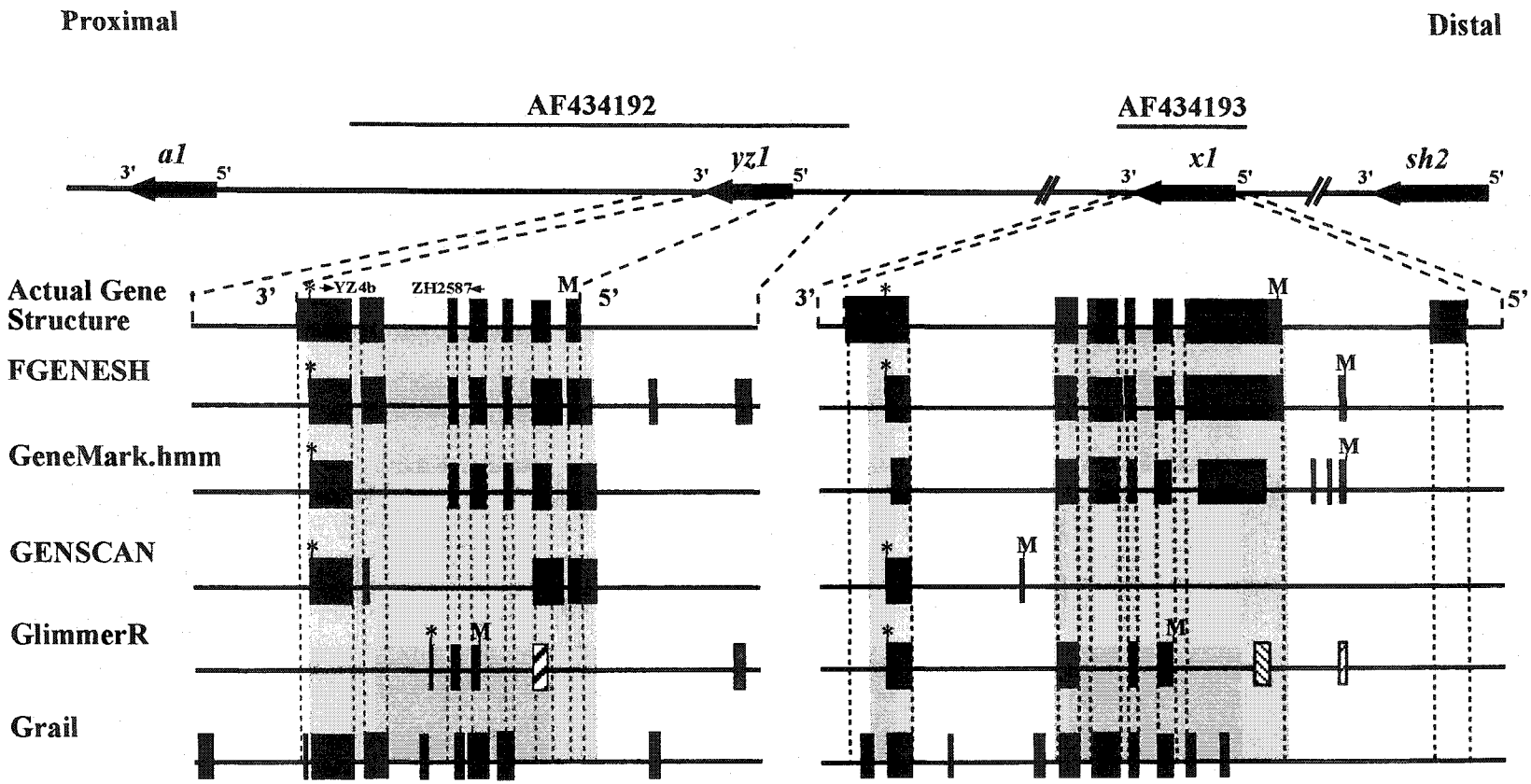


Figure 2. Yao *et al.*

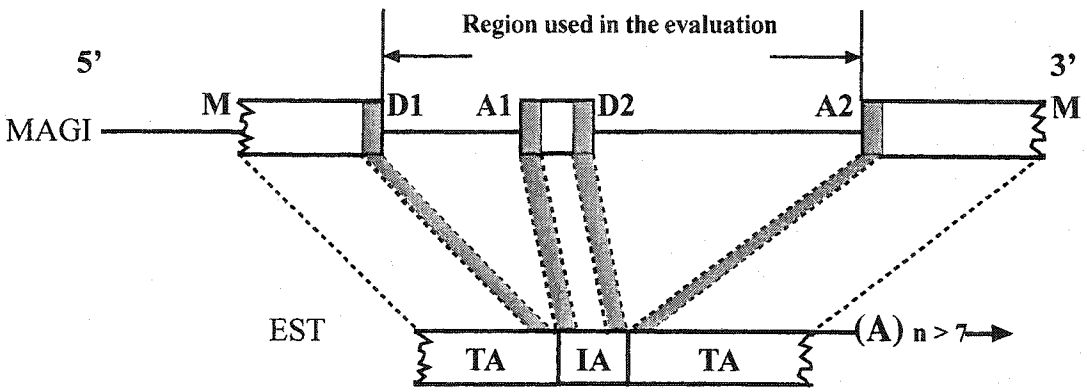


Figure 3. Yao *et al.*

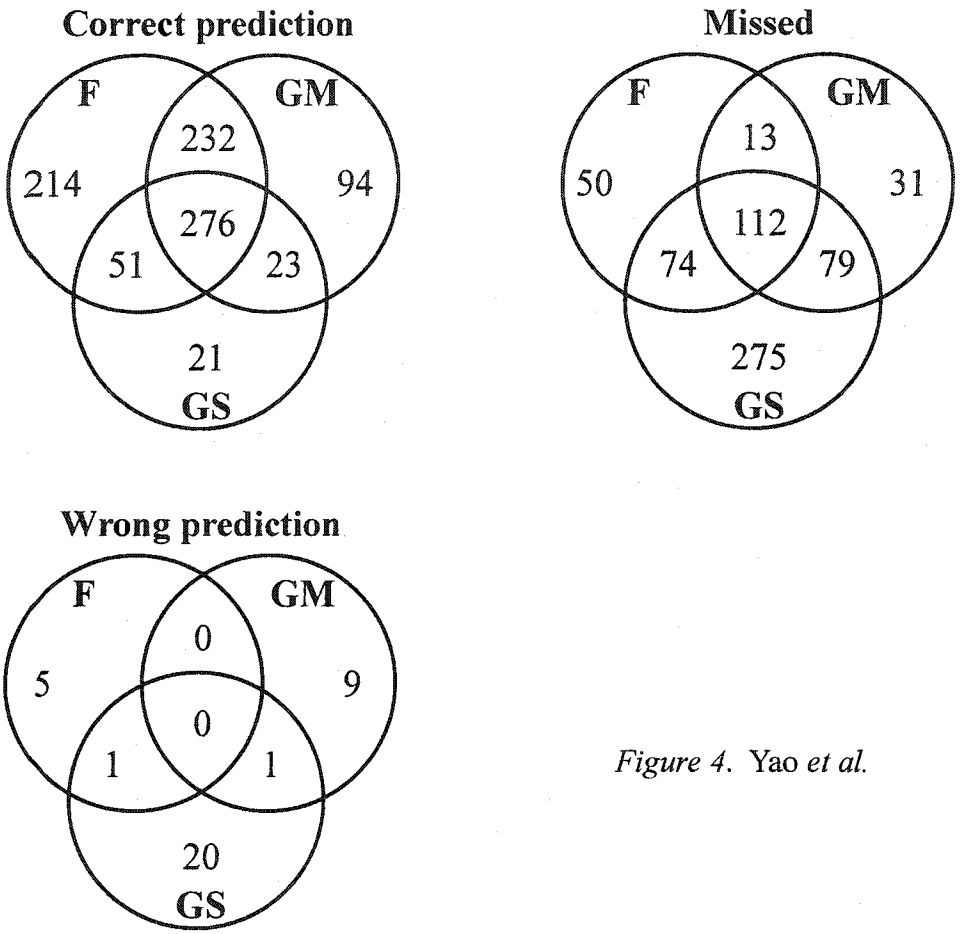


Figure 4. Yao *et al.*

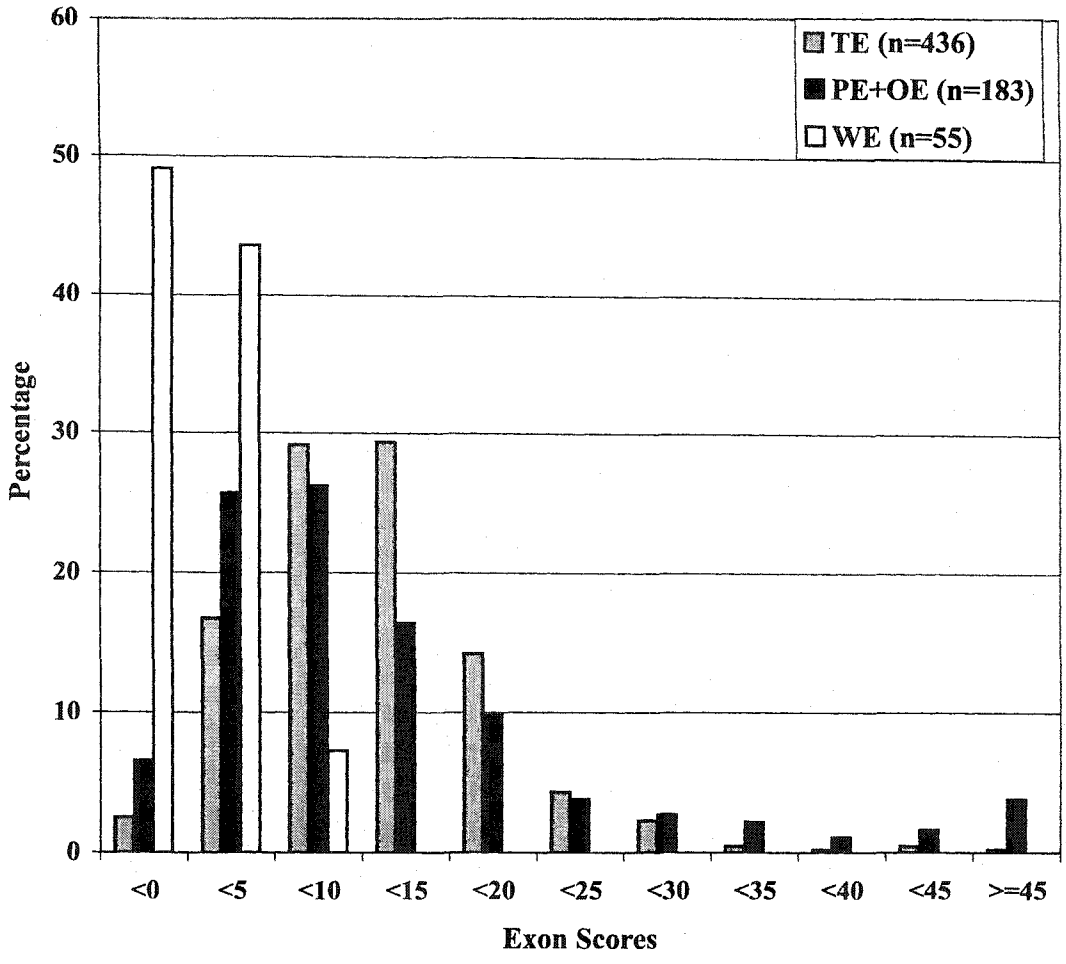


Figure 5. Yao *et al.*

Supplementary Materials

Methods

Methods used to create data set 2

Data set 2 were selected from the 114,173 contigs in the ISU MAGI 3.1b partial genome assembly (http://plantgenomics.iastate.edu/maize). This assembly was built from 879,523 *Zea mays* B73 genomic survey sequences (GSSs) that were postprocessed to significantly reduce sequencing errors within these data (Fu *et al.*, 2004). There are two additional improvements in this build as compared to the previous MAGI 2.3 assembly (Emrich *et al.*, 2004). First, Statistically Defined Repeats (SDRs) were obtained from a much larger collection of random maize sequences; this significantly improves repeat masking prior to clustering. Second, clone pairs are now used throughout the pipeline and both bridge gaps induced by masking and improve the assembly process. To efficiently identify correct gene models in these MAGIs, it is necessary to align the MAGIs with a B73 EST data set that consists of sequences with known gene orientations. It is also desirable to have available trace files to confirm alignments. For these reasons we used the approximately 32,000 3' B73 EST sequences (all of which are 3' reads) generated by Schnable Lab and that have been deposited in Genbank. Only those 30,356 ESTs that contained polyT prefixes of >7 bp (indicative of the presence of a polyA tail on the corresponding cDNA) were clustered using CAP3 (Huang and Madan, 1999); these polyT prefixes were masked prior to clustering. The clustering parameters were 98% identity, 60 bp overlap, 20 bp clipping range, and 5% overhang. In addition, the CAP3 program was not allowed to flip the input sequences during sequence assembly. This CAP3 analysis yielded 3,252 contigs and 16,202 singletons (these data are available from the authors upon request). To identify reliable genic regions

sequences of the MAGI contigs were then aligned with the clustered set of ESTs using the GeneSeqer program (http://bioinformatics.iastate.edu/cgi-bin/gs.cgi) (Usuka et al., 2000; Usuka and Brendel, 2000; Brendel *et al.*, 2004) with the default parameters for maize. GeneSeqer compares each orientation of each genomic sequence to each orientation of each EST. Only one of the four possible pairs of orientations can be biologically correct. All of the ESTs used in this study are known to be oriented 3' to 5' based on the cDNA cloning and sequencing strategies and confirmed by the presence of the polyT prefix. Hence, only the two pairs of orientations that involve the reverse complement of the EST sequence can be correct. Because in many EST collections the orientations are not known for certain, GeneSeqer was designed to align genomic sequences with ESTs having unknown orientations. This program allows a user to specify a particular orientation for the genomic sequences (using the –f and –r flags), but it does not allow a user to force GeneSeqer to consider only one orientation of the EST. The orientations of the ESTs used in this study are known. To avoid selection of an alignment that involves the wrong orientation of an EST, which can result in misalignments of genomic and EST sequences and the production of artifactual non-canonical splice sites, the set of ESTs was first compared to the entire set of genomic sequences using the -f flag, and then to the entire set of genomic sequences using the -r flag. The alignments obtained in the two experiments for each genomic sequence were then compared. For each genomic sequence, only that alignment in which the EST sequence had been flipped by GeneSeqer to the 5' to 3' orientation that was known to be correct was subjected to further analysis.

A MAGI contig and its corresponding GeneSeqer alignment positions and scores were parsed out if it contained at least two qualifying alignments (i.e., exons) to an EST.

Alignments were accepted only if they exhibited sequence similarities of at least 98%. As shown in Figure 3 in the manuscript, alignments between a genomic sequence and an EST can be either terminal or internal. Terminal alignments were required to be at least 50 bp. Although there was no minimum length requirement for internal alignments, they were accepted only if flanked by two qualifying alignments. In combination, these criteria resulted in the selection only of genomic sequences that contained at least one reliable intact intron. There are 4,405 qualified alignments corresponding to 2,326 MAGI contigs that passed this selection.

Among these MAGI contigs, 1,556 can be aligned with only one corresponding member of the clustered EST set; the remainder can be aligned with more than one EST member. The inclusion of genes that can be alternatively spliced has the potential to confound the evaluation of FGENESH. Hence, genomic sequences that can be aligned to more than one member of the clustered EST set were excluded from the analysis. Eighty-one out of the 1,556 MAGI contigs involve in the case where one EST can be aligned with multiple MAGI contigs and were not used in the FGENESH evaluation.

Hence, after selection candidates for data set 2 include 1,475 MAGI contigs that could be aligned to a single EST contig or singleton. Even among these MAGI contigs some splice sites could not be considered reliable and were therefore excluded from the analysis. For example, the endpoints of a terminal alignment reported by GeneSeqer can be incorrect even when associated with high similarity scores (data not shown). Hence, the endpoints of terminal alignments were excluded in the analysis (Figure 3). Also excluded were 41 MAGI contigs that contain canonical imperfect splice sites. An imperfect splice site is a site flanked by DNA sequence mismatches within 10 bp from the exon/intron junction in the alignment

between the genomic and EST sequences; a perfect splice site is the opposite. The rest MAGI contigs containing non-cononical splice sites were manually checked. Of these MAGI contigs, only those that have protein hits (E-value is greater or equal to 1e-15) to support the predicted gene orientation by GeneSeqer and that do not contain paired CT-AC sites (which is the reverse of the GT-AG canonical sites) were selected for further analysis. The orientations of the predicted genes in the MAGI contigs that failed this selection may not be correct, which could be caused by artifactual ESTs from the EST library. Hence, these MAGI contigs were removed from the candidate list. The selected MAGI contigs that contain non-canonical splice sites were checked for ambiguities in the exon/intron junctions. An ambiguous splice site means there is more than one way to interpret the splice junctions and in any of these interpretations the splice sites is non-canonical and perfect; such sites are therefore not suitable for FGENESH evaluation and the corresponding MAGI contigs were removed from the candidate list. The remaining MAGI contigs that contain imperfect or perfect non-canonical splice sites were further checked for possible sequence errors. First, these MAGI contigs were blasted against those GSSs that they were assembled from (the membership GSSs). If a non-canonical splice site is perfect and the sequence surrounding this site (200 bp upstream and downstream) in the MAGI contig is identical to the corresponding sequences of at least two corresponding membership GSSs, this non-canonical site is considered as reliable. Non-canonical splice sites in all other cases were checked in the regions surrounding these sites using trace files (NCBI TraceDB) of the corresponding membership GSSs and ESTs. Sequencing errors were corrected according to the trace files and the corresponding exon/intron junctions were double-checked for ambiguity. After such checking, one perfect non-canonical site turned out to be canonical after correcting sequence

error; another pair of non-canonical splice sites (one site is imperfect) could be interpreted as perfect canonical sites that are supported by both the GSS and EST sequences. Hence both cases were edited accordingly. Non-canonical splice sites that are still imperfect after manually checking were excluded from the analysis and the corresponding MAGI contigs were also removed from the candidate list. The remaining perfect unambiguous non-cononcial sites and all the perfect canonical sites (including the those converted from non-cononcial sites) were considered reliable and kept for evaluation. In summary, data set 2 consists of 1,353 selected MAGI contigs that contain at least one pair of reliable donor and acceptor sites flanking an intact intron (Figure 3). Alignments between the selected MAGI contigs in data set 2 and the corresponding EST sequences are available from the authors upon request. The statistical characteristics of data set 2 are shown in Figure 1.

Programs evaluated

The version of FGENESH (Salamov and Solovyev, 2000) evaluated in this study was trained on a data set that was created in 2001 and consists about 1,000 genes from rice, maize, wheat, barley and other monocots (Valery Sagitov, personal communication). The algorithm used by FGENESH is based on the Generalized Hidden Markov Model (GHMM). GeneMark.hmm was first developed to find genes in bacterial genomes (Lukashin and Borodovsky, 1998). It has since been trained for gene finding in both prokaryotic and eukaryotic genomes. GeneMark.hmm uses the GHMM as the framework of its algorithm and a ribosome binding site recognition algorithm is added to improve the prediction of the translation initiation codon. GENSCAN (version 1.0) (Burge and Karlin, 1997) evaluated in our study was trained on a small set of 41 maize genes that had been constructed by Dr.

Brendel (Kleffe *et al.*, 1996). The algorithm model of GENSCAN is also GHMM. It uses weight matrices, weight arrays and maximal dependence decomposition for signal (e.g., splice sites) recognition. GlimmerR is a special version of GlimmerM (Salzberg *et al.*, 1999). GlimmerR was trained specifically for rice gene discovery using a set of 172 complete genes and 133 partial genes (Yuan *et al.*, 2001). Its algorithm is based on the Interpolated Markov Model (IMM) to score the potential exons and the maximal dependence decomposition algorithm is used for splice site recognition. Grail (version 1.3) (Xu and Uberbacher, 1997) is based upon the neural network algorithm. The versions of Grail evaluated in this study was trained for *Arabidopsis* gene discovery.

References

Brendel, V., Xing, L. and Zhu, W. 2004 Gene structure prediction from consensus spliced alignment of multiple ESTs matching the same genomic locus. Bioinformatics 20: 1157-1169.

Burge, C. and Karlin, S. 1997. Prediction of complete gene structures in human genomic DNA. J. Mol. Biol. 268: 78-94.

Emrich, S.J., Aluru, S., Fu, Y., Wen, T.-J., Narayanan, M., Guo, L., Ashlock, D.A., Schnable, P.S. 2004. Astrategy for assembling the maize (*Zea mays* L.) genome. Bioinformatics 20: 140-147.

Fu, Y., Hsia, A.-P., Guo, L. and Schnable, P.S. 2004. Types and frequencies of sequencing errors in methyl-filtered and high C_ot maize genome survey sequences. Plant Physiology, in press.

Huang, X. and Madan A. 1999. CAP3: A DNA sequence assembly program. Genome Res. 9: 868-877.

Kleffe, J., Hermann, K., Vahrson, W., Wittig, B. and Brendel, V. 1996. Logitlinear models for the prediction of splice sites in plant pre-mRNA sequences. Nucleic Acids Res. 24: 4709-4718.

Lukashin, A.V. and Borodovsky, M. 1998. GeneMark.hmm: new solutions for gene finding. Nucleic Acids Res. 26: 1107-1115.

Salamov, A.A. and Solovyev, V.V. 2000. *Ab initio* gene finding in Drosophila genomic DNA. Genome Res. 10: 516-522.

Salzberg, S.L., Pertea, M., Delcher, A.L., Gardner, M.J. and Tettelin, H. 1999. Interpolated Markov models for eukaryotic gene finding. Genomics 59: 24-31.

Usuka, J. and Brendel, V. 2000. Gene structure prediction by spliced alignment of genomic DNA with protein sequences: increased accuracy by differential splice site scoring. J. Mol. Biol. 297: 1075–1085.

Usuka, J., Zhu, W. and Brendel, V. 2000. Optimal spliced alignment of homologous cDNA to a genomic DNA template. Bioinformatics 16: 203–211.

Xu, Y. and Uberbacher, E.C. 1997. Automated gene identification in large-scale genomic sequences. J. Comput. Biol. 4: 325-338.

Yuan, Q., Quackenbush, J., Sultana, R., Pertea, M., Salzberg, S.L. and Buell, C.R. 2001. Rice bioinformatics. analysis of rice sequence data and leveraging the data to other plant species. Plant Physiol. 125: 1166-1174.

CHAPTER 5. GENERAL CONCLUSIONS

Summary and Discussion

Meiotic recombination across the ~130-140-kb *a1-sh2* interval was characterized, aiming to answer the question "Why meiotic recombination occurs non-randomly in the maize genome?" The multigenic feature of this interval makes it suitable as a model system to study meiotic recombination in maize. Recombination breakpoints could be mapped to high resolution relative to genic and intergenic regions to compare their contributions to the recombination activity of the entire *a1-sh2* interval. In addition, genetic *cis*-effects on both intragenic and intergenic recombination in the *a1-sh2* interval could be characterized by using distinct maize and teosinte *a1 sh2* haplotypes.

Several conclusions were drawn based on the experimental data: 1) recombination breakpoints are not distributed uniformly across the *a1-sh2* interval, e.g., over 85% of the recombination breakpoints mapped to the *a1-yz1* subinterval that is only ~10% of the length of the *a1-sh2* interval; 2) these recombination breakpoints are clustered to hot spots that are genes and an apparently non-genic region. In addition, the *x1* gene as a whole is not a recombination hot spot. Hence, not all genes are recombination hot spots and not all hot spots are genes; 3) the retrotransposon fractions of the *a1-sh2* interval are recombinationally inert; 4) *cis*-genetic modifiers alter rate and distribution of recombination events across the *a1-sh2* interval. Consequently, not all genic and non-genic hot spots are conserved across different *A1 Sh2* haplotypes in a common genetic background; 5) sequence polymorphisms are generally negatively correlated with recombination in a region, i.e., recombination breakpoints tend to occur in less polymorphic segments of the region; 6) the frequency and distribution of sequence polymorphisms within a genic or intergenic region, however, are not sufficient to explain the non-random distribution of recombination breakpoints across a gene or an intergenic region. Region-specific chromatin structures and interactions between

adjacent regions (e.g., competing for initiation or resolution of recombination events) may also contribute to the regulation of recombination in genic and intergenic regions.

The conservation of the overall distribution of recombination events across the *a1-sh2* interval among distinct *A1 Sh2* haplotypes suggests that some features of the *a1-sh2* interval are not altered by sequence divergence among haplotypes. These features could be associated with the gross chromosome structures. Regulation of these structures could influence the recombination as suggested by studies in yeast and mammals (reviewed by PETES 2001; DE MASSY 2003). Genes are generally recombination hot spots in the maize genome. Intergenic regions are usually much less recombinationally active than genes. Even so, *cis*-modifiers can convert a genic hot spot to a non-hot spot and an intergenic cold spot to a hot spot in the *a1-sh2* interval. This *cis*-regulation of recombination is via local features of a region. These local features may include type, amount and distribution of sequence polymorphisms within this region, local chromatin structure and influence from adjacent regions. Genes that are favored in recombination may be so because they are generally much less polymorphic and usually have more attractive chromatin structures to recombination machinery than the intergenic regions.

At what step is recombination regulated to result in the non-random pattern of distribution, initiation or resolution? Studies of recombination in *S. cerevisiae*, mouse and human suggest that this regulation is via recombination initiation (reviewed by DE MASSY 2003; KAUPPI *et al.* 2004). Yet, without knowledge of distribution of recombination initiation sites (i.e., DSBs) relative to the recombination resolution sites in maize, no conclusions can be drawn based on our current experimental data. Recently, a PCR-based method developed in yeast to map meiosis-specific DSBs to high resolution has been successfully adapted in mouse (QIN *et al.* 2004). This method could be possibly optimized to

analyze meiotic DSBs in maize. The technical challenges are to enrich the meiosis-specific DSBs and to reduce the artifactual DSBs in the analysis.

References

PETES, T. D., 2001 Meiotic recombination hot spots and cold spots. Nat Rev Genet. **2:** 360-369.

DE MASSY, B., 2003 Distribution of meiotic recombination sites. Trends Genet. **19:** 514-522.

KAUPPI, L., A. J. JEFFREYS and S. Keeney, 2004 Where the crossovers are: recombination distributions in mammals. Nat. Rev. Genet. **5:** 413-424.

QIN, J., L. L. RICHARDSON, M. JASIN, M. A. HANDEL and N. ARNHEIM, 2004 Mouse strains with an active *H2-Ea* meiotic recombination hot spot exhibit increased levels of *H2-Ea*-specific DNA breaks in testicular germ cells. Mol. Cell. Biol. **24:** 1655-1666.

ACKNOWLEDGEMENTS

I would like to thank Dr. Patrick S. Schnable, my major professor, for his guidance in my graduate research and critical edition of this dissertation. Also I'd like to thank my committee members: Dr. Basil J. Nikolau, Dr. Charlotte Bronson, Dr. Thomas Peterson, and Dr. Volker Brendel for their helpful advice.

I appreciate encouragement and help from Ling Guo, Marna Yandeau, Jin Li, David S. Skibbe, Dr. Tsui-Jung Wen, Dr. An-Ping Hsia, Danette Bontrager, Marianne Smith and other members of Schnable lab.

Moreover, I would like to thank my parents and my husband, Dr. Weiwu Xie, for their love and support.